Encyclopedia of
Language and Education
*Series Editor:* Stephen May

SPRINGER
REFERENCE

Elana Shohamy · Iair G. Or
Stephen May *Editors*

# Language Testing and Assessment

*Third Edition*

Springer

# Encyclopedia of Language and Education

**Series Editor**
Stephen May
Faculty of Education and Social Work
The University of Auckland
Auckland, New Zealand

In this third, fully revised edition, the 10 volume *Encyclopedia of Language and Education* offers the newest developments, including an entirely new volume of research and scholarly content, essential to the field of language teaching and learning in the age of globalization. In the selection of topics and contributors, the Encyclopedia reflects the depth of disciplinary knowledge, breadth of interdisciplinary perspective, and diversity of sociogeographic experience in the language and education field. Throughout, there is an inclusion of contributions from non-English speaking and non-Western parts of the world, providing truly global coverage. Furthermore, the authors have sought to integrate these voices fully into the whole, rather than as special cases or international perspectives in separate sections. The Encyclopedia is a necessary reference set for every university and college library in the world that serves a faculty or school of education, as well as being highly relevant to the fields of applied and socio-linguistics. The publication of this work charts the further deepening and broadening of the field of language and education since the publication of the first edition of the Encyclopedia in 1997 and the second edition in 2008.

More information about this series at http://www.springer.com/series/15111

Elana Shohamy • Iair G. Or • Stephen May
Editors

# Language Testing and Assessment

## Third Edition

With 9 Figures and 3 Tables

Springer

*Editors*
Elana Shohamy
School of Education
Tel Aviv University
Tel Aviv, Israel

Iair G. Or
School of Education
Tel Aviv University
Tel Aviv, Israel

Stephen May
Faculty of Education and Social Work
The University of Auckland
Auckland, New Zealand

ISBN 978-3-319-02260-4 ISBN 978-3-319-02261-1 (eBook)
ISBN 978-3-319-02262-8 (print and electronic bundle)
DOI 10.1007/978-3-319-02261-1

Library of Congress Control Number: 2017934050

1st edition: © Kluwer Academic Publishers 1997
2nd edition: © Springer Science+Business Media LLC 2008
© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Editor in Chief's Introduction to the "Encyclopedia of Language and Education"

This is one of ten volumes of the *Encyclopedia of Language and Education* published by Springer. The *Encyclopedia* – now in this, its third edition – is undoubtedly the benchmark reference text in its field. It was first published in 1997 under the general editorship of the late David Corson and comprised eight volumes, each focused on a single, substantive topic in language and education. These included: language policy and political issues in education; literacy; oral discourse and education; second language education; bilingual education; knowledge about language; language testing and assessment; and research methods in language and education.

In his introductory remarks, David made the case for the timeliness of an overarching, state-of-the-art review of the language and education field. He argued that the publication of the *Encyclopedia* reflected both the internationalism and interdisciplinarity of those engaged in the academic analysis of language and education, confirmed the maturity and cohesion of the field, and highlighted the significance of the questions addressed within its remit. Contributors across the first edition's eight volumes came from every continent and from over 40 countries. This perhaps explains the subsequent impact and reach of that first edition – although no one (except, perhaps, the publisher!) quite predicted its extent. The *Encyclopedia* was awarded a Choice Outstanding Academic Title Award by the American Library Association and was read widely by scholars and students alike around the globe.

In 2008, the second edition of the *Encyclopedia* was published under the general editorship of Nancy Hornberger. It grew to ten volumes as Nancy continued to build upon the reach and influence of the *Encyclopedia*. A particular priority in the second edition was the continued expansion of contributing scholars from contexts outside of English-speaking and/or developed contexts, as well as the more effective thematic integration of their regional concerns across the *Encyclopedia* as a whole. The second edition also foregrounded key developments in the language and education field over the previous decade, introducing two new volumes on language socialization and language ecology.

This third edition continues both the legacy and significance of the previous editions of the *Encyclopedia*. A further decade on, it consolidates, reflects, and expands (upon) the key issues in the field of language education. As with its predecessors, it overviews in substantive contributions of approximately 5000

words each, the historical development, current developments and challenges, and future directions, of a wide range of topics in language and education. The geographical focus and location of its authors, all chosen as experts in their respective topic areas, also continues to expand, as the *Encyclopedia* aims to provide the most representative international overview of the field to date.

To this end, some additional changes have been made. The emergence over the last decade of "superdiversity" as a topic of major concern in sociolinguistics, applied linguistics, and language education is now a major thread across all volumes – exploring the implications for language and education of rapidly changing processes of migration and transmigration in this late capitalist, globalized world. This interest in superdiversity foregrounds the burgeoning and rapidly complexifying uses of language(s), along with their concomitant deconstruction and (re)modification, across the globe, particularly (but not exclusively) in large urban environments. The allied emergence of multilingualism as an essential area of study – challenging the long-held normative ascendancy of monolingualism in relation to language acquisition, use, teaching, and learning – is similarly highlighted throughout all ten volumes, as are their pedagogical consequences (most notably, perhaps, in relation to translanguaging). This "multilingual turn" is reflected, in particular, in changes in title to two existing volumes: *Bilingual and Multilingual Education* and *Language Awareness, Bilingualism and Multilingualism* (previously, *Bilingual Education* and *Language Awareness*, respectively).

As for the composition of the volumes, while ten volumes remain overall, the *Language Ecology* volume in the 2nd edition was not included in the current edition, although many of its chapter contributions have been reincorporated and/or reworked across other volumes, particularly in light of the more recent developments in superdiversity and multilingualism, as just outlined. (And, of course, the important contribution of the *Language Ecology* volume, with Angela Creese and the late Peter Martin as principal editors, remains available as part of the second edition.) Instead, this current edition has included a new volume on *Language, Education and Technology*, with Steven Thorne as principal editor. While widely discussed across the various volumes in the second edition, the prominence and rapidity of developments over the last decade in academic discussions that address technology, new media, virtual environments, and multimodality, along with their wider social and educational implications, simply demanded a dedicated volume.

And speaking of multimodality, a new, essential feature of the current edition of the *Encyclopedia* is its multiplatform format. You can access individual chapters from any volume electronically, you can read individual volumes electronically and/or in print, and, of course, for libraries, the ten volumes of the *Encyclopedia* still constitute an indispensible overarching electronic and/or print resource.

As you might expect, bringing together ten volumes and over 325 individual chapter contributions has been a monumental task, which began for me at least in 2013 when, at Nancy Hornberger's invitation, Springer first approached me about the Editor-in-Chief role. All that has been accomplished since would simply not have occurred, however, without support from a range of key sources. First, to Nancy Hornberger, who, having somehow convinced me to take on the role, graciously

agreed to be Consulting Editor for the third edition of the *Encyclopedia*, providing advice, guidance, and review support throughout.

The international and interdisciplinary strengths of the *Encyclopedia* continue to be foregrounded in the wider topic and review expertise of its editorial advisory board, with several members having had direct associations with previous editions of the *Encyclopedia* in various capacities. My thanks to Suresh Canagarajah, William Cope, Viv Edwards, Rainer Enrique Hamel, Eli Hinkel, Francis Hult, Nkonko Kamwangamalu, Gregory Kamwendo, Claire Kramsch, Constant Leung, Li Wei, Luis Enrique Lopez, Marilyn Martin-Jones, Bonny Norton, Tope Omoniyi, Alastair Pennycook, Bernard Spolsky, Lionel Wee, and Jane Zuengler for their academic and collegial support here.

The role of volume editor is, of course, a central one in shaping, updating, revising, and, in some cases, resituating specific topic areas. The third edition of the *Encyclopedia* is a mix of existing volume editors from the previous edition (Cenoz, Duff, King, Shohamy, Street, Van Deusen-Scholl), new principal volume editors (García, Kim, Lin, McCarty, Thorne, Wortham), and new coeditors (Lai, Or). As principal editor of *Language Policy and Political Issues in Education*, Teresa McCarty brings to the volume her longstanding interests in language policy, language education, and linguistic anthropology, arising from her work in Native American language education and Indigenous education internationally. For *Literacies and Language Education*, Brian Street brings a background in social and cultural anthropology, and critical literacy, drawing on his work in Britain, Iran, and around the globe. As principal editors of *Discourse and Education*, Stanton Wortham has research expertise in discourse analysis, linguistic anthropology, identity and learning, narrative self-construction, and the new Latino diaspora, while Deoksoon Kim's research has focused on language learning and literacy education, and instructional technology in second language learning and teacher education. For *Second and Foreign Language Education*, Nelleke Van Deusen-Scholl has academic interests in linguistics and sociolinguistics and has worked primarily in the Netherlands and the United States. As principal editors of *Bilingual and Multilingual Education*, Ofelia García and Angel Lin bring to the volume their internationally recognized expertise in bilingual and multilingual education, including their pioneering contributions to translanguaging, along with their own work in North America and Southeast Asia. Jasone Cenoz and Durk Gorter, principal editors of *Language Awareness, Bilingualism and Multilingualism*, bring to their volume their international expertise in language awareness, bilingual and multilingual education, linguistic landscape, and translanguaging, along with their work in the Basque Country and the Netherlands. Principal editor of *Language Testing and Assessment*, Elana Shohamy, is an applied linguist with interests in critical language policy, language testing and measurement, and linguistic landscape research, with her own work focused primarily on Israel and the United States. For *Language Socialization*, Patricia Duff has interests in applied linguistics and sociolinguistics and has worked primarily in North America, East Asia, and Central Europe. For *Language, Education and Technology*, Steven Thorne's research interests include second language acquisition, new media and online gaming environments, and

theoretical and empirical investigations of language, interactivity, and development, with his work focused primarily in the United States and Europe. And for *Research Methods in Language and Education*, principal editor, Kendall King, has research interests in sociolinguistics and educational linguistics, particularly with respect to Indigenous language education, with work in Ecuador, Sweden, and the United States. Finally, as Editor-in-Chief, I bring my interdisciplinary background in the sociology of language, sociolinguistics, applied linguistics, and educational linguistics, with particular interests in language policy, Indigenous language education, and bilingual education, along with my own work in New Zealand, North America, and the UK/Europe.

In addition to the above, my thanks go to Yi-Ju Lai, coeditor with Kendall King, and Iair G. Or, coeditor with Elana Shohamy. Also to Lincoln Dam, who as Editorial Assistant was an essential support to me as Editor-in-Chief and who worked closely with volume editors and Springer staff throughout the process to ensure both its timeliness and its smooth functioning (at least, to the degree possible, given the complexities involved in this multiyear project). And, of course, my thanks too to the approximately 400 chapter contributors, who have provided the substantive content across the ten volumes of the *Encyclopedia* and who hail from every continent in the world and from over 50 countries.

What this all indicates is that the *Encyclopedia* is, without doubt, not only a major academic endeavor, dependent on the academic expertise and goodwill of all its contributors, but also still demonstrably at the cutting edge of developments in the field of language and education. It is an essential reference for every university and college library around the world that serves a faculty or school of education and is an important allied reference for those working in applied linguistics and sociolinguistics. The *Encyclopedia* also continues to aim to speak to a prospective readership that is avowedly multinational and to do so as unambiguously as possible. Its ten volumes highlight its comprehensiveness, while the individual volumes provide the discrete, in-depth analysis necessary for exploring specific topic areas. These state-of-the-art volumes also thus offer highly authoritative course textbooks in the areas suggested by their titles.

This third edition of the *Encyclopedia of Language and Education* continues to showcase the central role of language as both vehicle and mediator of educational processes, along with the pedagogical implications therein. This is all the more important, given the rapid demographic and technological changes we face in this increasingly globalized world and, inevitably, by extension, in education. But the cutting-edge contributions within this *Encyclopedia* also, crucially, always situate these developments within their historical context, providing a necessary *diachronic* analytical framework with which to examine *critically* the language and education field. Maintaining this sense of historicity and critical reflexivity, while embracing the latest developments in our field, is indeed precisely what sets this *Encyclopedia* apart.

The University of Auckland
Auckland, New Zealand

Stephen May

# Volume Editors' Introduction to "Language Testing and Assessment"

This volume addresses the broad theme and specific topics associated with current thinking in the field of language testing and assessment. Interdisciplinary in their nature, language testing and assessment build on theories and definitions provided by linguistics, applied linguistics, language acquisition, and language teaching, as well as on the disciplines of testing, measurement, and evaluation. Language testing uses these disciplines as foundations for researching, theorizing, and constructing valid language tools for assessing and judging the quality of language. Language testing and assessment are always historically situated and conditioned, embedded in knowledge, beliefs, and ideologies about their goals and best practices. They also play an important role in education, policy, and society, and their educational and societal consequences cannot be ignored. The present volume therefore responds to the high demand for clear, reliable, and up-to-date information about language testing and assessment theories and practices, while keeping in sight the rich social contexts in which they function.

The main focus of this volume, which sets it apart from similar volumes and handbooks, is innovation. We wanted the volume to present state-of-the-art techniques, principles, insights, and methodologies for a new generation of practitioners, researchers, and experts in language testing and assessment. For this purpose, we selected a range of topics which, while providing a broad overview of the field, focuses on advances and breakthroughs of the past decade or so. As a consequence, many of the topics in this volume – such as multilingual assessment, the assessment of meaning, English as a lingua franca (ELF), the Common European Framework of Reference (CEFR), the Common Core policy in the USA, or critical testing – are covered for the first time in a volume of this sort by experts dedicated to them. Of the volume's 29 chapters, 15 are completely new, many of them covering aspects of language assessment that were not included in the second edition of this encyclopedia, published in 2008. In addition to that, we uniformly asked all the authors – both those contributing to the volume for the first time and those updating their contributions from the previous edition – to report about innovations, new research, or novel techniques in their area of expertise. Consequently, this third edition volume can be seen as groundbreaking, strongly emphasizing recent developments, as well as providing an outlook of the future of this dynamic field.

The field of language testing is traditionally viewed as consisting of two major components: one focusing on the "what," referring to the constructs that need to be assessed (also known as "the trait"), and the other component pertaining to the "how" (also known as "the method"), which addresses the specific procedures and strategies used for assessing the "what." Traditionally, "the trait" has been defined by the language testing field; these definitions have provided the essential elements for creating language tests. The "how," on the other hand, is derived mostly from the field of testing and assessment which has, over the years, developed a broad body of theories, research, techniques, and practices. Today, a crucial third component is added to the field, focusing on language assessment practices and the social consequences and implications of language testing and assessment. Language testers incorporate these three areas to create the discipline of language testing and assessment, a field which includes theories, research, and applications; it has its own research publications, conferences, and two major journals, *Language Testing* and *Language Assessment Quarterly*, where many of these studies appear.

An examination of the developments in the language testing and assessment field since the 1960s reveals that its theories and practices have always been closely related to definitions of language proficiency. Matching the "how" of testing with the "what" of language uncovers several periods in the development of the field, with each one instantiating different notions of language knowledge along with specific measurement procedures that go with them. Thus, discrete-point testing viewed language as consisting of lexical and structural items so that the language test of that era presented isolated items in objective testing procedures. In the integrative era, language tests tapped integrated and discoursal language; in the communicative era, tests aimed to replicate interactions among language users utilizing authentic oral and written texts; and in the performance testing era, language users were expected to perform tasks taken from "real life" contexts. Alternative assessment was a way of responding to the realization that language knowledge is a complex phenomenon, which no single procedure can be expected to capture. Assessing language knowledge therefore requires multiple and varied procedures that complement one another. While we have come to accept the centrality of the "what" to the "how" trajectory for the development of tests and assessment instruments, extensive work in the past two decades has pointed to a less overt but highly influential dynamic in another direction. This dynamic has to do with the pivotal roles that tests play in societies in shaping the definitions of language, in affecting learning and teaching, and in maintaining and creating social classes. This means that contemporary assessment research perceives as part of its obligations the need to examine the close relationship between methods and traits in broader contexts and to focus on how language tests interact with societal factors, given their enormous power. In other words, as language testers seek to develop and design methods and procedures for assessment (the "how") they become mindful not only of the emerging insights regarding the trait (the "what"), and its multiple facets and dimensions, but also of the societal role that language tests play, the power that they hold, and their central functions in education, politics, and society.

In terms of the interaction of society and language, it is evident that changes are currently occurring in the broader contexts and spaces in which language assessment takes place. It is increasingly realized nowadays that language assessment does not occur in homogeneous, uniform, and isolated contexts but, rather, in diverse, multilingual, and multicultural societies. This in turn poses new challenges and questions with regards to what it means to know language(s) in education and society. For example, different meanings of language knowledge may be associated with learning foreign languages, second languages, language by immersion, heritage languages, languages of immigrants arriving to new places with no knowledge of the new languages, multilingualism and translanguaging practices by those defined as "transnationals," and English as a lingua franca, for which language knowledge is different from the knowledge of other languages. As a consequence, the current focus on multilingualism, translanguaging, lingua franca, immigrants/refugees/asylum seekers, etc. has been incorporated in many of the chapters of this volume.

Similarly, the language of classrooms and schools may be different from that of the workplaces or communities where bi- or multilingual patterns are the norm. Each of these contexts may require different and varied theories of language knowledge and hence different definitions, applications, and methods of measuring these proficiencies. In other words, the languages currently used in different societies and in different contexts no longer represent uniform constructs, as these vary from one place to another, from one context to another, creating different language patterns, expectations, and goals, and often resulting in linguistic hybrids and fusions. Such dynamic linguistic phenomena pose challenges for language testers. What is the language (or languages) that needs to be assessed? Where can it be observed in the best ways? Is it different at home, in schools, in classrooms, and in the workplace? Should hybrids and fusions be assessed and how? Should multilingual proficiencies be assessed and how? Can levels of languages even be defined? How should language proficiency be reported and to whom? What is "good language"? Does such a term even apply? Who should decide how tests should be used? Do testers have an obligation to express their views about language and testing policy? What is the responsibility of testers to language learning and language use in classrooms and communities? How can ethical and professional attitudes in the field be maintained? These are some of the questions with which language testers are currently preoccupied. Language testers are not technicians that just invent better and more sophisticated testing tools. Rather, they are constantly in search for and concerned with the "what" and its complex meanings. Going beyond general testing, the unique aspect of language testing is that it is an integral part of a defined discipline, that of "Language." In this respect, language testers and the field of language testing and assessment are different from the field of general testing in that language testers are confined to a specific discipline and are therefore in constant need of asking such language-related questions as listed above in order to develop valid language assessment tools. Yet, even this list of questions is changing and context-dependent, since language today cannot be detached from multiple social, cultural, linguistic, and political dynamics.

The concern of language testers in the past two decades about the use of tests and their political, social, educational, and ethical dimensions has made the field even more complex and uncertain and in need of new discussions and debates. Elana Shohamy, the editor of this volume in the 2008 edition, stated that the era we are in could be described as the era of uncertainty, where questions are being raised about the meaning of language, along with the possibilities for measuring this complex and dynamic variable. While this statement still holds true, we may be experiencing times where some (complex, initial) answers and solutions for some of these questions are beginning to emerge. We are in an era where there is an ever more compelling need to ensure that these tests are reliable and valid, where validity includes the protection of the personal rights of others, as well as positive washback on learning by addressing the diverse communities in which the tests are used. Thus, the current era is not only concerned with a broader and more complex view of what it means to know a language, or with innovative methods of testing and assessment of complex constructs, but also with how these tests can be more inclusive, democratic, just, open, fair and equal, and less biased. Even within the use of traditional large-scale testing, the field is asking questions about test use: Why test? Who benefits, who loses? What is the impact on and consequences for definitions of language in relation to people, education, language policy, and society? Tests are no longer viewed as innocent tools, but rather as instruments that play central roles for people, education, and societies. Language testers, therefore, are asked to deal with and find solutions to broader issues: to examine the uses of tests in the complex multilingual and multicultural societies where they are used, not only as naïve measurement tools but also as powerful educational, societal, and political devices. This is the conceptual premise of this third edition volume of the *Encyclopedia of Language and Education* on Language and Assessment. It aims to cover (and uncover) the multiple versions and perspectives of the "what" of languages along with the multiple approaches developed for assessment of the "what," especially given the multiplicity of languages used by many diverse groups of learners in many different contexts. It aims to focus on the societal roles of language testers and their responsibility to be socially accountable and to ensure ethicality and professionalism. It also strives to show some of the emerging solutions and new directions that try to address these issues. A special focus is given in this volume to the multilingual and diverse contexts in which language testing and assessment are currently anchored and the difficult task of language testing and assessment in this complex day and age.

Accordingly, the first part of the volume addresses the "what" of language testing and assessment, looking into the constructs and domains of language assessment. Rather than dividing language into neat and clear-cut skills of reading, writing, speaking, and listening, it examines the "what" of language in the diverse contexts in which it is used. Instead of proposing one uniform way of defining the language construct, the chapters in Part 1 present language from multiple perspectives, which represent a variety of language activities. It begins with Lorena Llosa's chapter on the assessment of students' content knowledge and language proficiency, showing the complex, dynamic relations between content knowledge and language, critiquing

the traditional separation between the two and discussing recent attempts to integrate them in assessment. In the next chapter, Angela Scarino explores the position and role of culture in language assessment in times of increased globalization, multilinguality, and multiculturality. She argues that the construct of culture is and should be reconsidered to reflect complex realities, challenging established language assessment paradigms and raising ethical issues. James Purpura, in a novel contribution for such a volume, explores the construct of meaning and remaps the history of language testing through the lens of meaning-making. He shows that the focus since the 1980s on functional proficiency has been at the cost of meaning-making and propositional content and suggests various paths for assessing meaning. Rachel Brooks examines the changing language assessment practices and norms in the US government, as a large-scale example of language assessment at the workplace. Consisting of a wide range of departments, organizations, and aims, government activity greatly relies on high-stakes language testing, and some of its agencies are also involved in language testing development and research. Megan Smith and Charles Stansfield's chapter focuses on the language aptitude construct and the role of language aptitude tests in second language learning. The authors track the developments in the theory and practice of language aptitude measurement, as well as recent attempts to validate or find alternatives to the ways in which language aptitude is measured.

The concluding two chapters of the first part focus on recent challenges and innovations that represent two growing fields of language assessment. In their chapter on the assessment of multilingual competence, Alexis Augusto López, Sultan Turkan, and Danielle Guzman-Orth discuss the growing recognition, even by large testing authorities, that multilingual assessment tools are necessary for validly measuring the language knowledge of multilinguals in contexts of immigration or complex, globalized language realities. Although the field of multilingual assessment is still nascent, the authors present some of the early attempts that have already been made and discuss their importance and characteristics. Similarly, the chapter by Jennifer Jenkins on the assessment of English as a lingua franca (ELF) presents a field that seeks to answer the needs of globalized, transnational, "superdiverse" societies, in which English plays a major role as the shared language of non-native English speakers. Although no implementations of ELF tests and assessments have been developed so far, Jenkins outlines the goals, constructs, and limitations of such prospective tests, thereby proposing a novel outlook on how language testing can become more closely linked to the ways in which English is actually used as a second or foreign language. Together, these seven chapters provide multiple perspectives of the language constructs and assessment practices associated with them. As these chapters show, definitions of language cannot be detached from the diverse contexts in which they are used.

The second part of the volume addresses the methodological issues that language testers face when assessing the complex construct of language: that is, the "how." The chapters explore a wide variety of approaches and procedures for assessing language, each with its theoretical underpinnings and motivations and the issues it addresses. In the first chapter, Gillian Wigglesworth and Kellie Frost survey task and performance-based assessment, among the most popular alternative assessment tools

today, designed to measure learners' productive and receptive language skills through performances related to real world contexts. They discuss the value of certain performance tests, the extent to which they indeed represent "real life," and the recent trend of moving away from individual components of language proficiency to integrated tasks incorporating more than one skill. Staying within the context of alternative assessment, Janna Fox provides an overview of the various techniques, focusing on portfolio assessment, which has become the most pervasive approach. She discusses the usefulness of portfolios for both formative and summative assessment, as well as their claim for authenticity. Finally, she reviews the impact of newer technologies in the development of e-portfolios and other forms of digital learner records.

The implications of technology for language assessment are the topic of the next chapter, written by Carol Chapelle and Erik Voss, who begin their chapter with a historical overview of computer-assisted language testing, showing how technological advancements led to the development of computer-adaptive testing and natural language processing techniques. The authors discuss the potential influence of technology on test performance as part of the current and future challenges in the field. The chapter by Eunice Jang traces the cognitive processes involved in language assessment, looking into learner cognition and the way assessment tools should be devised to address various processes and their dynamic interplay with learners' multiple traits. Jang concludes the chapter by pointing to some future possibilities of harnessing technology to make assessment processes less intrusive. Glenn Fulcher provides a comprehensive description of the methods used for examining the quality of language via rating scales, standards, benchmarks, band levels, frameworks, and guidelines. He shows the advantages and disadvantages of these tools in terms of validity of progression, equivalence across languages, hierarchies, and misconceptions serving as criteria for language assessment. He stresses the fact that psychometrics has gone through major changes and has been replaced with a more pluralistic philosophical environment, in which consensus about language quality criteria no longer exists.

The chapter by Xiaoming Xi and Yasuyo Sawaki explores quantitative and qualitative methods of test validation, examining the evolution of validity theory and validation frameworks in general and argument-based validation in particular, and the issues associated with it. The authors also discuss the emergence of alternative validation approaches, constantly challenged by new concepts and constructs such as English as a lingua franca, new technologies, and new language learning frameworks. In continuation with the discussion of validation, Anne Lazaraton describes in her chapter the tensions between various approaches for validation and describes the increasingly popular qualitative approaches and techniques used for designing and evaluating performance tests. She surveys some of the key studies in this field, showing the merits of a mixed-methods approach, and discusses the main challenges faced by qualitative validation today. Concluding this section, Meg Malone's contribution focuses on training designed to increase language assessment literacy among teachers, principals, policy makers, and other agents. She reviews the major approaches in training, affected by changes in the educational, societal, and

philosophical contexts of testing. By analyzing textbooks for language assessment, she tracks the main developments in training and outlines some of the main issues, such as the scarcity of resources and lack of agreement between language testers and teachers regarding the main building blocks of language assessment literacy.

While the chapters of the second part highlight the practices and innovations in language assessment methods, from design to validation and training, the third part of this volume looks into language assessment as it is embedded in educational systems and contexts, where language assessment and especially tests are so widely used. It is in the educational system that tests and various assessment methods serve as major tools for: assessing language for learning and teaching, making decisions about programs, teachers and learners, and finally creating changes that lead to school reforms and bring intended and unintended washback in classrooms and schools. Matthew Poehner, Kristin Davin, and James Lantolf open this part with a chapter on dynamic assessment (DA), which is one of the most promising approaches to assessment in education. DA undertakes language assessment by applying Vygotsky's sociocultural theories, closely linking assessment and learning. The authors discuss the growing body of research in the field and emphasize the effectiveness of this approach with multiple populations, including immigrants, young learners, gifted learners, and learners with special needs. They conclude by discussing current studies on computerized administration of DA. Ofra Inbar-Lourie unravels the new concept of language assessment literacy (LAL) as an umbrella term for the knowledge, skills, and background that various participants in language assessment are expected to master. She explores the history of this concept and the challenges of arriving at an agreed upon set of skills or principles shared by the entire educational community. Looking into the future of this domain, she concludes that one of the most promising areas involves the creation of situated, differential LAL rather than a unified one.

The next five chapters are devoted to specific contexts of language assessment in education. Catherine Elder analyzes language assessment in the context of higher education, which is becoming a major site of Englishization and internationalization as well as language assessment expertise. Used for a wide variety of purposes, language assessment in higher education is often driven by powerful testing agencies, which in some cases limit the ability to develop local assessment policies for diverse student populations and for the introduction of new technologies. Beverly Baker and Gillian Wigglesworth delve into the Indigenous contexts of Australia and Canada – a research focus which is gaining recognition among researchers and policy makers. Against the backdrop of the historical mistreatment of Indigenous populations, both countries pay increased attention to language assessment as part of language revitalization and bilingual education efforts. The authors present some recent evidence showing that there is a growing acknowledgment of the importance of community participation in language assessment policies. Jamal Abedi looks into another intricate context of language assessment – that of using accommodations for learners with various disabilities or impairments, as well as for language learners in immigration contexts. Reviewing the extensive research conducted in the past two decades in the topic, he examines the effectivity and validity of accommodations for

language learners, mostly in the context of English language learners in the USA. He concludes with a set of principles regarding the need to limit the use of accommodations to the elimination of construct-irrelevant influences. Focusing on yet another language assessment context of expanding interest, Alison Bailey's chapter discusses young language learners (aged 3–11), who require a unique set of methods and techniques for assessing their language. Pointing to the different strategies of these kinds of tests compared with those used for adults, she explores the potential and limitations of the field, which is gaining major attention nowadays as it becomes ever more widely implemented. Constant Leung and Jo Lewkowicz complete this tour of language assessment contexts by surveying second or additional language assessment of linguistic minority students and in contexts where bi- or multilingualism is strongly encouraged, as in the European Union. They elucidate some of the constructs and recent developments, pointing at future directions which recognize the multiple linguistic repertoires and proficiencies of diverse populations and avoid the imposition of one language assessment standard on all.

Concluding the third part of the volume, Dina Tsagari and Liying Cheng delve into the study of the unavoidable washback, impact, and consequences assessment has on learning, teaching, and curriculum development. Tracking the long history of research into the impact and consequences of testing and distinguishing between two major strands of studies, they focus on recent studies, claiming that the complexity of these educational phenomena and the controversies surrounding them pose a serious challenge for any future study of these domains as well as for their interaction with notions of validity, fairness, and ethics in language assessment. Taken together, the chapters in Part 3 cover a wide range of topics related to broad issues of language assessment in education, especially amidst the changing realities of school demographics with regards to diverse populations and the role assessment can play in bringing about educational reform.

The fourth and final part of this volume puts language testing and assessment in a broader context, addressing the societal, political, professional, and ethical dimensions of assessments and tests. This topic has been a major concern in the language assessment field since the 1990s, and its importance is gaining broader recognition. Each of the six chapters in this section explores a different aspect of these dimensions. The section begins with a historical survey by Bernard Spolsky, in which the past, present, and future of the field are discussed, providing guidance and direction for the future. Spolsky surveys the advances in the field as well as the ample questions, contradictions, and uncertainties that need to be addressed in the future. He ends the chapter by stating that he remains skeptical about language testing, given the role of industrial test-makers in computerizing tests and in reducing multidimensional language profiles into uniform scales, and also given that educational systems continue to interpret test scores as if they are meaningful. At the same time, he expects the quality research that has been conducted in the field of language testing to continue, especially that which has been conducted in relation to the "nature" of language proficiency and the diverse approaches to assessing it in various social contexts. The chapter by Kate Menken illustrates how high-stakes language tests represent de facto language policies that affect schools and societies

and deliver direct messages about the significance and insignificance of certain languages and language instruction policies. Menken reviews the history of standardized testing and the detrimental impact of monolingual testing on education. She underlines the consequences of monolingual testing and proposes the adoption of multilingual assessment and translanguaging theory as a way to counter those problems, addressing immigrant and ELL populations.

The following chapter, on ethics, professionalism, rights, and codes, by the late Alan Davies, is included posthumously; we had the great honor of having him revise and update his contribution not long before his passing. Davies, who has written extensively on the ethical dimensions of tests and the professional aspects related to ethicality, addresses these issues by covering the developments in the language testing field, showing how the code of ethics and code of practice, developed by the language testing profession via the International Language Testing Association (ILTA), can lead to the more ethical use of tests, and questioning the effectiveness of this and similar courses of action. Davies warns against the use of ethical codes as face-saving devices, which, he argues, overlooks the real commitment to ethics that is instrumental for the profession itself, for its stakeholders, and for the rights of test-takers. He also proposes a model for the ethicality of tests for asylum seekers and the inappropriate use of tests by state authorities. This chapter is followed by two chapters that may illustrate some of the ethical complexities of language assessment, focusing on two major educational and societal contexts. First, Monica Barni and Luisa Salvati reflect on the uses and misuses of the Common European Framework (CEFR) for languages, originally designed to promote multilingualism and cultural diversity but eventually used by policy makers as a tool for the selection of migrant populations. Using the Italian situation as an example, the authors discuss the lack of reflection and consideration of the way the CEFR is used and the extent of its dangerous attraction for politicians and lawmakers, who tend to adopt it without considering the theory, know-how, and limitations of this tool from a professional point of view. Second, the chapter by Luis E. Poza and Guadalupe Valdés explores the recent history of English language assessment in the USA from the No Child Left Behind Act to the Common Core. The authors outline the tremendous impact of these two policies, which force schools and states to be constantly evaluated and particularly to develop or adopt new standards for English as a second language. The result has been the imposition of a standardizing testing-driven regime on English language learners (ELLs) who greatly vary in their levels of bilingualism and English-language proficiency. Poza and Valdés conclude by pointing at future directions that may mitigate some of the problems and improve the overall level of ESL, which is such a crucial component of education in the USA.

The concluding chapter of this volume, by Elana Shohamy, takes a critical look at testing by examining the critical issues arising from language testing in a variety of contexts. She discusses the critical language testing (CLT) research agenda proposed by her and other authors in the past two decades, focusing on the power of tests and the ways it can and should be addressed. By going back to many of the contributions in this volume, Shohamy points at various directions in which current research in the language assessment domain can tackle the issues created by the often detrimental

effects of language testing, suggesting constructive and positive forms of language assessment, enhancing equality and justice in this domain, and encompassing new definitions of language that are more pertinent to our times.

The editors would like to thank each and every author of these chapters, which together make up a most valuable contribution to current thinking in the field of language testing and applied linguistics. The authors selected to write these chapters are among the most distinguished scholars and leaders in the field of language testing and assessment internationally. The chapters herein reveal that the language testing field is dynamic, thriving, and vital. It is clear from these chapters that the field of language testing raises deep, important questions and does not overlook problems, difficulties, contradictions, malpractices, and new societal realities and needs. While viewed by some as a technical field, this volume convincingly demonstrates that language testing and assessment is, above all, a scholarly and intellectual field that touches the essence of languages in their deepest meanings. The need to get engaged in testing and assessment forces testers to face these issues head-on and attempt to deliberate on creative and thoughtful solutions which benefit society and are professional and ethically responsible.

Tel Aviv

Elana Shohamy
Iair G. Or

# Acknowledgment

Thanks to the following for their key editorial support:
Consulting Editor: Nancy Hornberger
Editorial Assistant: Lincoln Dam

# Contents

# About the Editors

**Elana Shohamy** is a Professor of Multilingual Education at Tel Aviv University where she teaches and researches various sociolinguistic issues related to coexistence and rights in multilingual societies within the frameworks of critical language testing, language policy, migration, and linguistic landscape. Her current work in language testing focuses on multilingual assessment. Elana authored *The Power of Tests: A Critical Perspective on the Uses of Language Tests* (2001), *Language Policy: Hidden agendas and new approaches* (2006), and she edited various books on the above topics. She was the editor of the 2008 edition of the present volume on language testing and assessment of the *Encyclopedia of Language Education* (Springer). Elana served as the editor of the journal *Language Policy* (2007–2015) and the founder and current editor of the new journal *Linguistic Landscape*. Professor Shohamy was granted the 2010 ILTA Lifetime Achievement Award in Cambridge LTRC for her work on critical language testing.

**Iair G. Or** is a PhD candidate at the Tel Aviv University School of Education, Multilingual Education Program, researching Hebrew and Arabic language policy in Israel. His book *Creating a Style for a Generation: The Beliefs and Ideologies of Hebrew Language Planners* was published in Hebrew in 2016 and contains a historical analysis of the linguistic ideologies in the Hebrew language revival discourse in Ottoman and British-ruled Palestine. Or's published articles focused on critical comparative analysis of Arabic and Hebrew in the Israeli educational system and matriculation exams, as well as the implementation of international standards such as the CEFR. Or received the 2013 AAAL Graduate Student Award for his study on language ideologies in the planning of Modern Hebrew. He teaches courses on language testing and alternative assessment at the MA TESOL and Multilingual Education programs at Tel Aviv University and Kibbutzim College, Tel Aviv.

**Stephen May** is Professor of Education in Te Puna Wānanga (School of Māori and Indigenous Education) in the Faculty of Education and Social Work, University of Auckland, New Zealand. He is an international authority on language rights, language policy, bilingualism and bilingual education, and critical multicultural approaches to education and, to date, has published 15 books and over 90 articles and chapters in these areas. His key books include *The Multilingual Turn* (2014), *Language and Minority Rights* (2nd edition, 2012), and, with Christine Sleeter, *Critical Multiculturalism: Theory and Praxis* (2010). In addition to being Editor-in-Chief of the 3rd edition of the *Encyclopedia of Language and Education*, he is a Founding Editor of the interdisciplinary journal, *Ethnicities*, and was from 2005 to 2015 Associate Editor of *Language Policy*. He is also a Fellow of the American Educational Research Association (AERA). His homepage is http://www.education.auckland.ac.nz/uoa/stephen-may.

# Advisory Board

**Suresh Canagarajah** Pennsylvania State University, University Park, PA, USA

**William Cope** Common Ground Publishing, Champaign, IL, USA

**Jim Cummins** University of Toronto, Toronto, ON, Canada

**Viv Edwards** University of Reading, Reading, UK

**Eli Hinkel** Seattle University, Seattle, WA, USA

**Francis Hult** Lund University, Lund, Sweden

**Nkonko Kamwangamalu** Howard University, Washington, DC, USA

**Gregory Kamwendo** University of Kwazulu-Natal, Durban, South Africa

**Claire Kramsch** University of California, Berkeley, CA, USA

**Rainer Enrique Hamel** Universidad Autónoma Metropolitana, Mexico City, Mexico

**Constant Leung** King's College, London, UK

**Li Wei** University College London, London, UK

**Luis Enrique Lopez** Universidad Mayor de San Simón, Cochabamba, Bolivia

**Marilyn Martin-Jones** University of Birmingham Edgbaston, Birmingham, UK

**Bonny Norton** University of British Columbia, Vancouver, BC, Canada

**Tope Ominiyi** Roehampton University, London, UK

**Alastair Pennycook** University of Technology, Sydney, Australia

**Bernard Spolsky** Bar-Ilan University, Ramat Gan, Israel

**Lionel Wee** National University of Singapore, Singapore, Singapore

**Jane Zuengler** University of Wisconsin, Madison, WI, USA

# External Reviewers

## Language Testing and Assessment

Ofra Inbar-Lourie
Darla Deardorff
Pardee Lowe
Kate Menken
Durk Gorter
Andy Kirkpatrick
Anne Lazaraton
Aek Phakiti
Constant Leung
Meg Malone
Elana Spector-Cohen
Richard Hill
Sultan Turkan

# Contributors

**Jamal Abedi** School of Education, University of California, Davis, CA, USA

**Alison L. Bailey** Human Development and Psychology Division, Department of Education, University of California, Los Angeles, CA, USA

**Beverly Baker** Official Languages and Bilingualism Institute, University of Ottawa, Ottawa, ON, Canada

**Monica Barni** University for Foreigners of Siena, Siena, Italy

**Rachel L. Brooks** Federal Bureau of Investigation, Washington, DC, USA

**Carol A. Chapelle** Applied Linguistics, Department of English, Iowa State University, Ames, IA, USA

**Liying Cheng** Faculty of Education, Queen's University, Kingston, ON, Canada

**Alan Davies** University of Edinburgh, Edinburgh, UK

**Kristin J. Davin** Foreign and Second Language Education, School of Education, Loyola University Chicago, Chicago, IL, USA

**Catherine Elder** School of Languages and Linguistics, University of Melbourne, Parkville, VIC, Australia

**Janna Fox** School of Linguistics and Language Studies, Carleton University, Ottawa, ON, Canada

**Kellie Frost** Language Testing Research Centre, School of Languages and Linguistics, University of Melbourne, Parkville, VIC, Australia

**Glenn Fulcher** English Department, School of Arts, University of Leicester, Leicester, UK

**Danielle Guzman-Orth** Educational Testing Service, Princeton, NJ, USA

**Ofra Inbar-Lourie** The School of Education, The Program for Multilingual Education, Tel Aviv University, Tel Aviv, Israel

**Eunice Eunhee Jang** Department of Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada

**Jennifer Jenkins** Department of Modern Languages, University of Southampton, Southampton, UK

**James P. Lantolf** Language Acquisition and Applied Linguistics, Department of Applied Linguistics, Center for Language Acquisition, CALPER, The Pennsylvania State University, University Park, PA, USA

**Anne Lazaraton** Department of Writing Studies, University of Minnesota, Minneapolis, MN, USA

**Constant Leung** Centre for Language, Discourse and Communication, School of Education, Communication and Society, King's College London, London, UK

**Jo Lewkowicz** University Council for the Certification of Language Proficiency, University of Warsaw, Warszawa, Poland

**Lorena Llosa** Multilingual Multicultural Studies, Department of Teaching and Learning, New York University, New York, NY, USA

**Alexis A. Lopez** Educational Testing Service, Princeton, NJ, USA

**Margaret E. Malone** Assessment and Evaluation Language Resource Center, Department of Linguistics, Georgetown University, Washington, DC, USA

**Kate Menken** Department of Linguistics, Queens College, City University of New York, Flushing, NY, USA

Research Institute for the Study of Language in an Urban Society, Graduate Center, City University of New York, New York, NY, USA

**Matthew E. Poehner** World Languages Education and Applied Linguistics, Department of Curriculum and Instruction, The Pennsylvania State University, University Park, PA, USA

**Luis E. Poza** School of Education and Human Development, University of Colorado Denver, Denver, CO, USA

**James Enos Purpura** Teachers College, Columbia University, New York, NY, USA

**Luisa Salvati** University for Foreigners of Siena, Siena, Italy

**Yasuyo Sawaki** Faculty of Education and Integrated Arts and Sciences, Waseda University, Tokyo, Japan

**Angela Scarino** Research Centre for Languages and Cultures, School of Communication, International Studies and Languages, Adelaide, SA, Australia

**Elana Shohamy** School of Education, Tel Aviv University, Tel Aviv, Israel

**Megan Smith** Department of English, Mississippi State University, East Lansing, MS, USA

**Bernard Spolsky** Bar-Ilan University, Ramat Gan, Israel

**Charles W. Stansfield** Language Learning and Testing Foundation, Rockville, MD, USA

**Dina Tsagari** Department of English Studies, University of Cyprus, Nicosia, Cyprus

**Sultan Turkan** Educational Testing Service, Princeton, NJ, USA

**Guadalupe Valdés** Race, Inequality, Language and Education (RILE) Program, Graduate School of Education, Stanford University, Stanford, CA, USA

**Erik Voss** NU Global, Northeastern University, Boston, MA, USA

**Gillian Wigglesworth** Research Unit for Indigenous Language, ARC Centre of Excellence for the Dynamics of Language, Faculty of Arts, University of Melbourne, Parkville, VIC, Australia

**Xiaoming Xi** New Product Development, Educational Testing Service, Princeton, NJ, USA

# Part I

# Assessing Language Domains

# Assessing Students' Content Knowledge and Language Proficiency

Lorena Llosa

**Abstract**

The relationship between language proficiency and content knowledge in assessment is a complicated one. From the perspective of content assessment, language has typically been considered a source of construct-irrelevant variance. From the perspective of language assessment, content has also been considered a potential source of construct-irrelevant variance. However, regardless of the purpose for assessment, both content knowledge and language proficiency are engaged to some extent. This chapter explores how the relationship between these two constructs has been conceptualized in the field of language assessment.

**Keywords**

Language assessment • Content assessment • English language learners

## Contents



L. Llosa (✉)
Multilingual Multicultural Studies, Department of Teaching and Learning, New York University, New York, NY, USA
e-mail: lorena.llosa@nyu.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_33

## Introduction

The relationship between language proficiency and content knowledge in assessment has always been a complicated one. From the perspective of content assessment, language has typically been considered a source of construct-irrelevant variance – variance in scores that is not related to the construct being assessed. From the perspective of language assessment, content (also referred to as topical knowledge or background knowledge) has also been considered a potential source of construct-irrelevant variance. Thus for the purpose of assessment, language proficiency and content knowledge have traditionally been viewed as separate and distinct constructs. The language ability models that have informed the constructs of most language assessments (e.g., Bachman and Palmer 1996) included topical knowledge as a category of language use, but one that was separate from language knowledge and strategic competence.

Regardless of the purpose of an assessment – either to assess a test taker's language proficiency or their content knowledge in a particular area – these two constructs cannot be so easily disentangled. Any assessment of content will involve language, and any assessment of language that will be useful for making inferences about a test taker's ability to use language in a context outside the test itself will involve some content or topical knowledge. Therefore the nature of the content-language link and the role it plays in construct definitions when assessing learners of a second or additional language has become an important concern in the field of assessment.

The need to better understand the relationship between language proficiency and content knowledge emerged initially in the context of bilingual education and the content-based instruction movement in the 1990s (Byrnes 2008). Since then, the need has only increased. As a result of immigration and globalization, a sizable proportion of students in schools and universities are learning content in a second or additional language. In the USA, for example, almost 10% of school-aged children are classified as English language learners (ELLs) (NCES 2015). Also, the workforce continues to become more global, and many workers carry out their profession in a second or additional language. In many parts of the world, English's role as a lingua franca has meant that students often learn content in English in addition to their first language. The popularity of the content and language integrated learning (CLIL) movement in Europe, which involves the teaching and learning of content through a foreign language or lingua franca (typically English), is another example of a context in which language and content interact (Dalton-Puffer 2011). Finally, over the past couple of decades, there has been an increase in the number of English-medium universities (EMUs) and programs in places where English is a second or foreign language. English-medium education is most prevalent in Europe but is quickly expanding throughout the world (Wilkinson et al. 2006). Although important work on the relationship between language and content has been conducted in relation to CLIL and EMUs, the primary concerns in terms of assessment have been the language assessment policies and practices affecting the students and the faculty

in these programs. The focus has not yet shifted to the integration of language and content in assessment (see Hofmannová et al. (2008) for emerging work on assessment that integrates language and content in a CLIL course). Wilkinson et al. (2006) assert that "the fact that education takes place through a language that is not the students' mother tongue (and, in many cases, not that of the educators either) seems to have little influence on the assessment processes" (p. 30). They explain that "the typical approach would be to apply assessment processes that are virtually the same as would be applied in the mother tongue context" (pp. 29–30). Given that the focus of this chapter is on the relationship between language and content in construct definitions in assessment, the remainder of the chapter will focus on areas of research where this relationship has been explicitly explored.

## Early Developments

Content-based instruction changed the landscape of language teaching by shifting the focus from communication in general to content as a context for language learning (Brinton et al. 1989). It is in the context of content-based instruction and bilingual education programs that concerns about the relationship between content and language began to be explicitly articulated (Byrnes 2008). As Short (1993) explains, in this context English learners needed to be involved in "regular curricula before they have fully mastered the English language" since "there simply is no time to delay academic instruction until these students have developed high levels of English language proficiency if they are to stay in school, succeed in their classes, and graduate with a high school diploma" (p. 628) – a claim still valid and relevant today for students around the world who are in school systems where they learn content in a second or additional language. Short strongly promotes the use of alternative assessments over standardized tests for assessing students in integrated language and content courses and programs, including the use of skill checklists and reading/writing inventories, anecdotal records and teacher observations, student self-evaluations, portfolios, performance-based tasks, essay writing, oral reports, and interviews. Even though she acknowledges "some overlap will occur between the language and content," she argues that when it comes to assessment, "it is more advisable to focus on a single objective, be it content or language specific" (pp. 634–35).

## Major Contributions

Major contributions to our understanding of the relationship between language proficiency and content knowledge in assessment emerged from the following areas of research: (1) language for specific purposes (LSP) testing and (2) content and language assessment of ELLs in schools.

## Language for Specific Purpose Testing

The complicated relationship between content and language has long been acknowledged in the field of languages for specific purposes (LSP). Davies (2001), for example, argued that "LSP testing cannot be about testing for subject specific knowledge. It must be about testing the ability to manipulate language functions appropriately in a wide variety of ways" (p. 143). Douglas (2005), however, stated that the defining characteristic of LSP assessment is "a willingness, indeed a necessity, to include nonlinguistic elements in defining the construct to be measured" (p. 866). In fact, he argued that LSP testing "is defined by the nature of the construct to be measured, which includes both specific purpose language and background knowledge" (p. 866). One way in which background or content knowledge has been taken into account in LSP assessment is by incorporating "indigenous assessment criteria" (Jacoby and McNamara 1999), that is, assessment criteria derived from the target language use domain.

A recent example of a study that identifies the indigenous criteria that underlie professional judgments of communication in the context of the health professions is that of Elder et al. (2012). The rationale for their investigation, as for much of the work on LSP assessment, is that "if LSP tests are to act as proxies for the demands of communication faced by candidates entering the workforce, then the judgments of such professionals should not be ignored" (p. 409). In their study, they asked several health professionals to provide feedback on video recordings of trainee-patient interactions from the Occupational English Test, a specific-purpose English language test used in Australia for overseas-trained health professionals. Performances on this test are assessed using primarily linguistic criteria, including intelligibility, fluency, appropriateness of language, resources of grammar and expression, and overall communicative effectiveness.

They found that the health professionals in their study rarely mentioned language skills in their feedback about the performances they observed. The authors hypothesize that the health professionals' lack of attention to language skills may be "because they give priority to clinical matters, because they feel that commenting on such features is beyond their competence, because they are blind to them (i.e., they lack the skills to make a linguistic diagnosis) or, more radically, because such features are irrelevant to what counts in clinical communication in their view" (p. 416). Elder et al. (2012) speculate that it may be that the candidates evaluated were already above a certain threshold of language proficiency that allowed the health professionals to focus on the clinical aspects of the performance. Uncovering the precise reasons for why the health professionals did not attend to language skills would be an important next step to better understand the role of content and language in this particular context.

Focusing on another LSP context, aviation English, Emery (2014) reflects on developments in the field in the last 30 years. He argues that the major change has been "the acceptance that it is neither possible nor desirable to separate language knowledge from subject matter knowledge" (p. 213). Nonetheless, he notes that "the extent and nature of the relationship between subject matter knowledge and

performance on language tests and the threat this represents to the validity of test scores" continues to be a key issue in LSP testing. He explains, however, that in the case of aviation English where those assessed are trained and licensed professional pilots and air traffic controllers with high level of expertise in their field, "the question of whether it is possible or even desirable to separate subject matter knowledge from language knowledge is perhaps less relevant." (Emery 2014, p. 210).

In fact, LSP testing in general often focuses on adults with high levels of expertise in a particular field. For this population, the challenge might simply be identifying the minimum threshold level of proficiency needed for communication. It may be that beyond that level of proficiency, language no longer plays an important role. The challenge for the field of LSP then would be identifying what that threshold is. Content and language assessment in schools, however, present different challenges in that students are developing both their language proficiency and their content knowledge at the same time.

## Content and Language Assessment of ELLs in Schools

A greater focus on testing and accountability in many countries around the world has resulted in more assessments of students, including those learning in a second or additional language. In the USA, for example, No Child Left Behind (2001) required that all students including ELLs had to be assessed in the content areas of English language arts, mathematics, and science. The legislation also required that ELLs' language proficiency had to be assessed annually. The need to assess all students in the content areas and the fact that a large proportion of students in schools are ELLs prompted discussions about the challenge of assessing ELLs' content knowledge in English. Similarly, the need to annually assess ELLs' language proficiency prompted discussions about the most appropriate and useful ways to do so. At the heart of these discussions was the content-language link.

**Content-language link in content assessments.** The main challenge in assessing ELLs in the content areas in English had been the score interpretation. Does the score on a content assessment represent the student's content knowledge or does it represent their ability to read, understand, and respond to questions in English? Abedi (2004) argues that language is a source of construct-irrelevant variance when assessing ELLs in the content areas and that scores from these assessments are not meaningful indicators of students' content knowledge. This perspective is supported by correlational studies that have found a relationship between the presence of complex linguistic features in test items and greater relative difficulty of the items for ELLs (e.g., Wolf and Leon 2009). Accommodations, modifications made to the assessment or the assessment administration, were introduced as a way to provide ELLs an opportunity to demonstrate their mastery of the content (Abedi et al. 2004). The assumption underlying accommodations is that language and content are separate constructs and that students will be able to demonstrate their content knowledge if their language ability does not get in the way.

However, research on the effectiveness of accommodations meant to reduce the linguistic load of test items has yielded mixed results, raising questions about this assumption (Kieffer et al. 2009, 2012). Outcomes of this research have led to a consensus on the need to better understand the language-content link. At minimum, it is important to distinguish "between language abilities central to the academic skills being measured and language demands of the test that are not relevant to the skills and abilities being measured" (Kieffer et al. 2012, p. 3). Avenia-Tapper and Llosa (2015) propose an approach for making distinctions between construct-relevant and construct-irrelevant language in content assessments. Drawing from systemic functional linguistics, they argue that certain complex linguistic features are a component of content area mastery, and thus, complex linguistic features cannot be considered construct-irrelevant on the basis of their complexity alone. Instead, the strong presence or absence of the linguistic features in the domain to which the test should generalize (e.g., grade-level science talk and text) is a better criterion for judging the relevance of a given linguistic feature. This approach would prevent assessment developers from eliminating complex structures that may be critical to the content area, thus guarding against the possibility of creating accommodated tests that suffer from construct underrepresentation, which could in turn cause negative washback for ELLs.

**Content-language link in English language proficiency (ELP) assessments.** Research on English language proficiency tests developed prior to NCLB uncovered that the language assessed by these tests did not align with the types of academic language that students needed to succeed in school (e.g., Stevens et al. 2000). Work was carried out to define and operationalize the construct of academic language proficiency by investigating empirically the kinds of English required of K–12 ELLs (Bailey and Butler 2003). Various categorizations of academic language emerged, describing it in terms of its lexical, grammatical, and textual characteristics (Bailey 2007).

An important shift in thinking about academic English proficiency was reflected in the ELP standards that emerged in 2004, which differed markedly in their conceptualization of English proficiency from most existing ELP standards. The existing ELP standards focused on language as communication and tended to be closely aligned to English language arts standards. The ELP standards, developed by the WIDA consortium (2004, 2007) and then augmented and adopted by TESOL (2006), were designed to link ELP to social and instructional language and to four content areas – language arts, mathematics, science, and social studies. Research on the ACCESS for ELLs, the ELP assessment designed to measure students' mastery of the WIDA standards, revealed that even though the assessment taps primarily into a language construct, content is assessed to some extent as well, especially at the higher levels of English proficiency. Romhild et al. (2011) identified "domain-general" and "domain-specific" linguistic knowledge factors underlying the structure of various forms (by grade and level of language proficiency) of this ELP assessment. Domain-general linguistic knowledge referred to academic language common to various content areas, whereas domain-specific knowledge referred to academic language specific to a particular content area. They found that the domain-

general factor was stronger in most forms of the test, but in forms assessing higher levels of English proficiency, the domain-specific factor was stronger than the domain-general factor. In other words, in assessments focused on students' mastery of ELP standards that link language proficiency to the content areas, it became difficult to disentangle language proficiency from content knowledge at higher levels of language proficiency.

## Work in Progress

Work in progress in the area of K–12 assessment in the USA has the potential to inform and transform the way the relationship between language proficiency and content knowledge is envisioned and how these constructs will be assessed in the future. A new wave of standards has emerged through the Common Core State Standards (CCSS) for English language arts and literacy in history, social studies, science, and technical subjects, the CCSS in mathematics (Common Core State Standards Initiative 2010a, b), and the Next Generation Science Standards (NGSS Lead States 2013). A major feature of these new standards is an emphasis on literacy and *practices* that are language and discourse rich. For example, "engage in argument from evidence" is one of the NGSS practices, "comprehend as well as critique, value evidence" are included in the CCSS for English language arts, and "construct viable arguments and critique the reasoning of others" is in the CCSS in mathematics (Stage et al. 2013).

These standards represent a major shift in the way content is defined, taught, and assessed. For the past decade, ELP standards have moved toward the content areas, whereas now standards in the content areas are moving toward language. Responding to the demands of these new standards for all students and using these demands as opportunities to help ELLs will require new ways of thinking about the relationship between language and content learning (Valdés et al. 2014). The *Understanding Language* Initiative at Stanford University (http://ell.stanford.edu) has led the effort to support ELLs in meeting new content standards, adopting a view of language as action that focuses on the essential role of language in learning academic content. For example, as part of the *Understanding Language* Initiative, in the area of science education for ELLs, Lee et al. (2013) suggest "(a) a shift away from both content-based language instruction and the sheltered model to a focus on language-in-use environments and (b) a shift away from 'teaching' discrete language skills to a focus on supporting language development by providing appropriate contexts and experiences" (p. 228). They introduce a conceptual framework that illustrates how the science and engineering practices in the NGSS can be unpacked into the types of language and discourse needed to instantiate these practices.

Two consortia are developing ELP assessments that are aligned to the new content standards. WIDA revised its standards and its assessment, ACCESS for ELLs. The revised standards still link language proficiency to social and instructional language and the four content areas, but are more explicit about how academic language is conceptualized by outlining specific features at the word/phrase,

sentence, and discourse level. At the word/phrase level, the focus is on vocabulary usage; at the sentence level, the focus is on language forms and conventions; and at the discourse level, the focus is on linguistic complexity. The second consortium, the English Language Proficiency Assessment for the twenty-first century (ELPA21) consortium, has developed ELP standards as a foundation to their assessment system informed by the work of the *Understanding Language Initiative*. ELPA21 specified ten standards that focus on form (e.g., vocabulary, grammar, and discourse specific to particular content areas) and function (e.g., what students do with language to accomplish content-specific tasks).

## Problems and Difficulties

The lack of empirical research about the development of and relationship between content knowledge and language proficiency remains a major challenge. Byrnes (2008) explains that "because content knowledge in an L2 learning environment is even more a developmental matter than is the case for native language instruction, content assessment would benefit from principles that identify how content and language abilities develop simultaneously in language learning" (p. 45). In the context of K-12 assessment in schools specifically, there is a lack of research on the relationship between (academic) language development and content instruction for all students, not just ELLs (Frantz et al. 2014). This lack of empirical research and the fact that both language proficiency and content knowledge develop across grades makes it particularly difficult to establish boundaries between these constructs. These boundaries are important as long as there is a need or mandate to assess language proficiency and content knowledge separately as is the case in the USA and in many other countries. Another reason why it might be important to locate these boundaries is to be able to use assessment information diagnostically. It may be helpful for educators to be able to identify sources of students' difficulty in accomplishing a task, whether it be language, content, or both.

## Future Directions

New task types and advances in technology may allow us to better understand the content-language link and develop assessments that assess content and language in an integrated way and at the same time allow for some separation of the two constructs. Integrated tasks, tasks that assess more than one language skill, have been developed in the past several years in response to increased awareness of the complexity of language use and the importance of context. The TOEFL iBT, for example, includes integrated tasks that require students to read a passage, listen to a lecture, and respond in writing. Integrated tasks are believed to be more representative of actual language use and thus allow for score-based interpretations that can be generalized to a particular target language use domain. A similar rationale could be

applied to justify the development of integrated tasks of language proficiency and content knowledge.

Given the focus on language and literacy skills in the content areas in the new standards, new content assessments will need to embrace these broader definitions of content and engage students' rich language use. Thus, a separate assessment of ELP may not be needed; it may be possible to assess language proficiency and content knowledge within the same assessment (Bailey and Wolf 2012). One technology-based innovation that would lend itself to integrated assessments of language and content are scenario-based assessments. These types of assessments are specifically designed to assess learners' integrated skills in a purposeful, interactive, and strategic manner. Scenario-based assessments have been used primarily to assess reading skills (Sabatini et al. 2014), but their use for ELP assessment is already being explored. In the content area of science, simulation-based assessments have been developed for both high-stakes summative assessment and classroom formative assessment (e.g., Quellmalz et al. 2012). Simulation-based assessments allow students to demonstrate their science knowledge as well as their ability to engage in scientific practices (e.g., predicting, observing, explaining findings, arguing from evidence). It may be possible to add a language dimension to these simulations so that language skills, which are already elicited as part of the assessment of science practices, are assessed alongside science content.

Finally, another innovation in assessment that would make integrated assessments of content and language particularly useful for instructional and diagnostic purposes is the use of scaffolds embedded in technology-enhanced assessments. Wolf and Lopez (2014) have examined the impact of including scaffolds in a scenario-based assessment of young ELLs' language proficiency. Their assessment includes speaking tasks with scaffolding questions: Students first retell a story independently, then answer scaffolding questions, and then retell the story for a second time. They found that students were more successful in retelling the story after responding to the scaffolding questions and that low-performing students on the task were at least able to complete the scaffolding questions. They concluded that "the incorporation of scaffolding into assessment has the potential to improve the measurement of EL students' language proficiency and also provide useful information for teachers' instruction." Both content and language scaffolds could be incorporated into technology-enhanced and scenario- or simulation-based assessments. In fact, simulation-based assessments already have the capability to provide scaffolds and immediate feedback and coaching related to the science knowledge and inquiry practices being assessed by the simulations (Quellmalz et al. 2012). These assessments have also experimented with accommodations for ELLs and student with disabilities, including audio recordings of text, screen magnification, and segmentation to support reentry at the beginning of a task to allow for extended time (Quellmalz et al. 2012). Much more refined language scaffolds could be added to these simulations to allow ELLs of different levels of proficiency to engage with the tasks and demonstrate both their content knowledge and their language proficiency. Scaffolds could be informed by current work on learning progressions in the content areas (see NGGS Lead States 2013 as an example) and in specific areas of language

development (Bailey and Heritage 2011–15). These types of innovative, technology-enhanced, simulation-based, scaffolded assessments could be used both to assess and promote learning and also as a means to investigate the developmental nature of content and language learning for ELLs.

## Cross-References

- ▶ Assessing Meaning
- ▶ Dynamic Assessment
- ▶ Language Assessment in Higher Education
- ▶ Language Assessment Literacy

## Related Articles in the Encyclopedia of Language and Education,

Masaki Kobayashi, Sandra Zappa-Hollman, Patricia Duff: Academic Discourse Socialization. In Volume: Language Socialization

Mary R. Lea: Academic Literacies in Theory and Practice. In Volume: Literacies and Language Education

Tarja Nikula: CLIL: A European Approach to Bilingual Education. In Volume: Second and Foreign Language Education

Fredricka L. Stoller, Shannon Fitzsimmons-Doolan: Content-Based Instruction. In Volume: Second and Foreign Language Education

## References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4–14.

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Avenia-Tapper, B., & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment, 20*, 95–111.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, A. L. (2007). *The language demands of school: Putting academic English to the test*. New Haven: CT Yale University Press.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Technical Report No. 611). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Bailey, A. L., & Heritage, M. (2011–15). Dynamic Language Learning Progressions Project. www.dllp.org

Bailey, A. L., & Wolf, M. K. (2012). The challenge of assessing language proficiency *aligned* to the Common Core State Standards and some possible solutions. Retrieved from http://ell.stanford.edu/papers/policy

Brinton, D., Snow, M. A., & Wesche, M. B. (1989). *Content-based second language instruction*. New York: Newbury House.

Byrnes, H. (2008). Assessing content and language. In E. Shohamy and N. H. Hornberger (Eds.), *Encyclopedia of language and education, 2nd edition, volume 7: Language testing and assessment* (pp. 37–52). New York: Springer Science+Business Media.

Common Core State Standards Initiative. (2010a). Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects. Retrieved from http://www.corestandards.org/ELA-Literacy

Common Core State Standards Initiative. (2010b). Common Core State Standards for mathematics. Retrieved from http://www.corestandards.org/Math

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics, 31*, 182–204.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*(2), 133–147.

Douglas, D. (2005). Testing languages for specific purposes. In E. Hinkel (Ed.), *The handbook of research in second language teaching and learning* (pp. 857–868). Mahwah: Erlbaum.

Elder, C., Pill, J., Wookward-Kron, R., McNamara, T., Manias, E., Webb, G., & McColl, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *Tesol Quarterly, 46*(2), 409–419.

Emery, H. J. (2014). Developments in LSP testing 30 years on? The case of aviation English. *Language Assessment Quarterly, 11*(2), 198–215.

Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the United States. *Language Assessment Quarterly, 11*, 432–457.

Hofmannová, M., Novotná, J., & Pípalová, R. (2008). Assessment approaches to teaching mathematics in English as a foreign language. *International CLIL Research Journal, 1*(1), 20–35.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*, 213–241.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201.

Kieffer, M. J., Rivera, M., & Francis, D. J. (2012). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments. 2012 update*. Portsmouth: RMC Research Corporation, Center.

Lead States, N. G. S. S. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to next generation science standards and with implications for common core state standards for English language arts and mathematics. *Educational Researcher, 42*(4), 223–233.

National Center for Education Statistics. (2015). The Condition of Education 2015 (NCES 2015–144). Retrieved from https://nces.ed.gov/programs/coe/indicator_cgf.asp.

No Child Left Behind Act. (2001). Pub. L. No. 107–110, § 115, Stat. 1425(2002).

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

Romhild, A., Kenyon, D., & McGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly, 8*(3), 213–228. doi:10.1080/15434303.2011.558146.

Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice, 29*(1), 36–43.

Short, D. J. (1993). Assessing integrated language and content instruction. *TESOL Quarterly, 27*, 627–656.

Stage, E. K., Asturias, H., Cheuk, T., Daro, P. A., & Hampton, S. B. (2013). Opportunities and challenges in next generation standards. *Science, 340*, 276–277.

Stevens, R., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs*. Final Deliverable to OERI, Contract No. R305B60002. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Teachers of English to Speakers of Other Languages. (2006). *PreK-12 english language proficiency standards*. Alexandria, Virginia: TESOL.

Valdés, G., Kibler, A., & Walqui, A. (2014). *Changes in the expertise of ESL professionals: Knowledge and action in an era of new standards*. Alexandria: TESOL International Association.

Wilkinson, R., Zegers, V., & van Leeuwen, C. (2006). *Bridging the assessment gap in English-Medium higher education*. Maastricht University Language Centre.

Wolf, M., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3), 139–159.

Wolf, M., & Lopez, A. (2014). *The use of scaffolding strategies in the assessment of English Learners*. Paper presented at the Roundtable on Learning-Oriented Assessment in Language Classrooms and Large-Scale Assessment Contexts in New York.

# Culture and Language Assessment

Angela Scarino

**Abstract**

This first entry on culture and language assessment is written at a time of much reconsideration of the major constructs in language/s learning and language assessment. This is in response at least partly to the increasingly complex reality of multilinguality and multiculturality in our contemporary world. Culture is one of these constructs and is considered in its interrelationship with language and learning. It is because of this reconsideration that the discussion in this chapter is focused on scoping the conceptual landscape and signaling emerging rather than established lines of research. The discussion encompasses (a) the assessment of culture in the learning of languages, including recent interest in assessing intercultural practices and capabilities, and (b) the role of culture (and language), or its influence, on the assessment of learning where multiple languages are in play. The discussion considers the place of culture in conceptualizing the communicative competence and understandings of the role of culture in all learning. Developments related to the assessment of intercultural practices and capabilities in foreign language learning are described, as well as multilingual (and multicultural) assessment approaches. The assessment of capabilities beyond the linguistic poses major challenges to traditional conceptualizations and elicitation and judgment practices of assessment. This is because what is being assessed is the linguistic and cultural situatedness of students of language/s as they communicate and learn across linguistic and cultural systems. This challenges the traditional assessment paradigm and also raises important ethical issues. This conceptual and practical stretch can only extend thinking about educational assessment.

A. Scarino (✉)
Research Centre for Languages and Cultures, School of Communication, International Studies and Languages, Adelaide, SA, Australia
e-mail: angela.scarino@unisa.edu.au

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_3

**Keywords**

Assessing culture • Culture in assessment • Multilingual assessment • Intercultural capabilities • Translanguaging • Hermeneutic approaches

## Contents



## Introduction

It is timely to consider culture and language assessment, as culture is a dimension that has been undergoing major reconsideration in language/s learning in the past decade (e.g., see Byrnes 2010), and yet it is underrepresented in the language assessment literature.

The discussion in this chapter will consider mainly the assessment of culture/s in the learning of language/s, including the recent interest in assessing intercultural practices and capabilities in language/s learning. This refers to how "cultural knowing" or "cultural/intercultural understanding" is assessed in the context of learning language/s. The discussion will also consider, to a lesser extent, the role of culture, or its influence, on the assessment of learning in environments where multiple languages are in play and where students are or are becoming multilingual. This aspect highlights that the process of assessing learning (of language/s or other disciplines) is itself both a cultural (and linguistic) act and that culture/s (and language/s) come into play in learning and in the assessment of learning. This is because students are linguistically and culturally situated in the linguistic and cultural systems of their primary socialization. In developing their learning and in assessment, they draw upon their own dynamic histories of experiences of knowing, being, and communicating and their own frameworks of values and dispositions. In discussing both aspects, the focus will remain specifically on education and educational assessment.

The discussion takes as a starting point the move in language/s teaching and learning, away from a monolingual and national paradigm (with the one language equals one culture equation) toward a multilingual paradigm. [For a detailed discussion, see the guest-edited volume of The Modern Language Journal by Claire Kramsch (MLJ, 98, 1, 2014), "Teaching foreign languages in the era of globalisation."] It is this move that gives greater prominence to the interplay of multiple languages and with these multiple cultures, in all learning and therefore in assessment.

In a recent 25-year review of culture in the learning of foreign languages, Byram (2014), one of the most prolific writers on the role of culture in language teaching, learning, and assessment, observed that "the question of assessment remains insufficiently developed"(p. 209). Atkinson (1999) reflected on how little direct attention is given to the notion of culture in TESOL, even though "ESL teachers face it in everything they do" (p. 625). Block (2003), discussing the social turn in second language acquisition (SLA), raised questions about "a cultural turn" for SLA research. He specifically noted the difficulty involved in conceptualizing the relationship between language and culture, but also the promising work in pragmatics and in learner identity as research areas that take culture into account. Although Shohamy (2011) did not specifically address culture, she drew attention to an important dimension of the discussion when arguing for assessing "multilingual competencies" in an assessment field that continues to view language as a monolingual, homogenous, and often still native-like construct (p. 419). I add that the monolingual bias that Shohamy described in language assessment extends to it also being a *monocultural* bias.

These reflections signal some of the efforts to reconsider and expand the constructs in language teaching, learning, and assessment "beyond lingualism" (Block 2014) to include dimensions such as subjectivity and identity.

## Early Developments

Culture comes into play in the diverse contexts of language learning and assessment, both as a dimension of the *substance* of learning (e.g., in the learning of foreign languages) and as the *medium* for learning language/s and other areas of learning (e.g., in the learning of ESL/EAL). In considering culture as substance, it is necessary to consider the relationship between language and culture. In considering culture as medium, it is necessary to consider the relationship between language and culture and learning.

### Language and Culture

The integral relationship between language/s and culture/s has long been considered from diverse disciplinary perspectives, including linguistics, anthropology, sociolinguistics, and applied linguistics (Whorf 1940/1956; Sapir 1962; Geertz 2000; Gumperz and Levinson 1996; Kramsch 2004). In the diverse contexts of language/s

learning, this interrelationship is understood and foregrounded in different ways. In foreign language teaching and learning, culture has been understood traditionally as factual knowledge or as a form of "content" of language learning, with literature and other aesthetic forms as rich expressions of particular culture/s. In this sense culture is understood as observable products or artifacts, associated with a particular social group. It has also been understood as ways of life, behaviors, and actions of a social group where the language/s is used. Both of these understandings present a static view of culture that removes variability and personal agency within the national group. A more recent perspective is an understanding of culture as social norms and practices, created through the use of language (see Byrnes 2010). Such practices, however, are removed from the cultural identity of the learner as a participant in language learning. In ESL/EAL, where the major goal is to prepare students for learning in English across diverse disciplines, the interrelationship between language and culture has been backgrounded in order to focus on subject matter learning across the curriculum.

A useful starting point for a consideration of culture and language assessment is how it has been represented in the construct of "communicative competence." This is because it is the conceptualization of the construct that guides elicitation, judging, and validation in the assessment cycle (Scarino 2010). In the conceptualization of "linguistic competence," where the focus was on the linguistic system itself, there was an absence of any attention to culture or to language users as participants in the linguistic and cultural system. Canale and Swain's (1980) framework comprised grammatical competence (vocabulary and rules of grammar), sociolinguistic competence (conventions of use), discourse competence (cohesion and coherence of texts), and strategic competence (compensating for limited resources in using language). This modeling highlighted the social, interactive nature of language use and the crucial role of context. The sociolinguistic interest here was with how the social context affects choices within the linguistic system. Halliday's theoretical work is instructive in this regard.

Halliday (1999) used the theoretical constructs "context of situation"' and "context of culture" to explain what is entailed in an exchange of meanings in communication. In Halliday's terms, these two constructs do not refer to "culture" in the sense of lifestyles, beliefs, and value systems of a language community (e.g., as in traditional foreign language learning) but rather as a system of meanings. He makes clear that the two constructs are not two different things, but rather that they are the same thing seen from two different depths of observation. The "situation" provides the context for particular instances of language use, and, as such, it is an instance of the larger system, which is referred to as "culture." For Halliday, culture is in the very grammar that participants use in exchange.

Bachman and Palmer (1996, 2010) built on the Canale and Swain model by identifying "knowledge" in the mind of the user, which can be drawn upon in communication. They identified (a) organizational knowledge, that is, grammatical knowledge and textual knowledge, and (b) pragmatic knowledge, that is, functional and sociolinguistic knowledge. Pragmatic knowledge is understood as objective

knowledge that is necessary for selecting language appropriately for use in particular social situations. As such, it represented a static view of the context of situation and of participants in that context. Although this is recognized as the most developed model of "communicative ability" for the purposes of assessment, it has been criticized because of its individualistic view of social interaction (McNamara and Roever 2006) and because context is not sufficiently taken into account (Chalhoub–Deville 2003; see also Bachman 2007). In the extensive discussion about context in defining the construct of communicative competence in language assessment, the context has been understood essentially as the context of situation, with little explicit attention to the context of culture.

The applied linguist who has most extensively theorized culture in (foreign) language learning is Claire Kramsch. In her 1986 critique of the proficiency movement as an oversimplification of human interaction, Kramsch extended the construct from communicative to "interactional competence." She highlighted at the same time that this interaction takes place within "a cross-cultural framework" (p. 367) and that successful interaction necessitates the construction of a shared internal context or "sphere of intersubjectivity" (p. 367). This understanding of culture foreshadowed her extensive discussion of context and culture in language teaching (Kramsch 1993) and her subsequent theorization of culture as "symbolic competence" (Kramsch 2006), which I consider below (see section "Major Contributions").

## Language, Culture, and Learning

Language and culture are integral to learning. Halliday (1993) highlighted learning itself as a process of meaning-making when he wrote:

> When children learn a language, they are not simply engaging in one kind of learning among many; rather, they are learning the foundation of learning itself. The distinctive characteristic of human learning is that it is a process of meaning making – a semiotic process. (p. 94)

It is through language, in the context of situation and the context of culture, that students and teachers, in their diversity, interact to exchange knowledge, ideas, explanations, and elaborations and make sense of and exchange meaning in learning. In the learning interaction, this meaning is mediated through the lenses of the language/s and culture/s of participants' primary socialization.

All learning, therefore, is essentially a linguistic and cultural activity. It is formed through individual learners' prior knowledge, histories, and linguistic and cultural situatedness. It is the learner's situatedness and the cultural framing of learning that shapes the interpretation and exchange of meanings in learning and, by extension, in the assessment of learning. This understanding is in line with cultural views of learning in education. Gutierrez and Rogoff (2003) described learning as emerging from participating in practices, based on students' linguistic and cultural–historical repertoires. Lee (2008) also discussed "the centrality of culture to . . . learning and

development" (p. 267). This understanding of the relationship between language, culture, and learning is related to the sociocultural family of theories of language learning, in which the role of culture at times remains implicit. This understanding of learning as a linguistic, social, and cultural act of meaning-making becomes important in assessment. Shohamy (2011) expressed concern with the differential performance of immigrant students, depending on whether they are assessed in the language of their primary socialization or in the language of education in their new locality. The meanings that students make and represent in learning and assessment necessarily originate in the linguistic, cultural, experiential, and historical knowledge context to which they belong. It is this relationship that underlies Shohamy's argument for multilingual assessment (see section "Future Directions" below).

## Major Contributions

Major contributions to the consideration of culture and language assessment have been advanced in relation to ongoing conceptualizations of the construct of communicative competence, including toward "intercultural competence," the assessment of intercultural practices and capabilities, and multilingual approaches to assessment.

### Ongoing Conceptualization of Communicative Competence Toward "Intercultural Competence"

In more recent work, Kramsch has expanded further the constructs of communicative competence and interactional competence to what she has termed "symbolic competence" (Kramsch 2006). In her conceptualization, knowledge of and engagement with the systems of culture associated with language provide the basis for understanding the ways in which users of the language establish shared meanings, how they communicate shared ideas and values, and how they understand the world. Language constitutes and reflects the social and cultural reality that is called context. Symbolic competence foregrounds meaning-making not only as an informational exchange but as a process of exchange of cultural meaning, including its interpretive and discursive symbolic dimensions. It entails using language to negotiate and exchange meanings in context, both reciprocally with others and in individual reflection on the nature of the exchanges. Context is not fixed or given but created in interaction through the intentions, assumptions, and expectations of participants. Kramsch foregrounded not only such exchange *within* a language but also *across* languages and cultures in multilingual and multicultural contexts, and it is in this way that she elaborated foreign language learning as an intercultural endeavor that develops "intercultural competence."

## Assessing Intercultural Competence

Perhaps because her conceptualization of culture and the intercultural in language/s learning is the most elaborated and complex, Kramsch (2009) questioned whether or not it can be assessed. She stated:

> [S]ymbolic competence based on discourse would be less a collection of. . . stable knowledges and more a savviness i.e., a combination of knowledge, experience and judgment. . . Trying to test symbolic competence with the structuralist tools employed by schools. . . is bound to miss the mark. Instead, symbolic competence should be seen as the educational horizon against which to measure all learners' achievements. (p. 118)

This may well be the case within traditional testing paradigms, but it has been suggested that possibilities may be available within alternative assessment paradigms (Scarino 2010) and assessment purposes that are educational.

In considering assessment in the context of intercultural language learning, a major distinction needs to be drawn between the consideration of "intercultural understanding" in general education, where language is not foregrounded (Bennett 1986), and in language/s education, where language use and language learning are the focus.

The extensive efforts to model intercultural competence began with Byram and Zarate (1994) and Byram (1997), working under the auspices of the Council of Europe. Their conceptualization was based on a set of knowledge, skills, and dispositions called *savoirs*: *savoir apprendre*, *savoir comprendre/faire*, *savoir être*, and *savoir s'engager*. In line with the council's orientation, it was focused on an objectives-setting approach, which was analytic rather than holistic, and on defining levels of intercultural competence. Although these *savoirs* captured broad educational dimensions such as *savoir être* (knowing how to be) and *savoir s'engager* (knowing how to engage politically), the original modeling did not sufficiently foreground communication. Byram (1997) subsequently modeled "intercultural communicative competence," incorporating the set of dimensions of the model of Canale and Swain (1981), discussed above, with the set of *savoirs* that defined intercultural competence. As with all modeling, however, the relationship among these sets of dimensions was not explained. Risager (2007) included further dimensions, which she described as "linguacultural competence," resources, and transnational cooperation, thereby highlighting the multilingual (and multicultural) nature of communication. Sercu (2004) considered the inclusion of a "metacognitive dimension" that focuses on students monitoring their learning. Although this is a valuable dimension, Sercu did not specify that the reflective work should be focused on exploring the linguistic and cultural situatedness of participants involved in communication and learning to communicate interculturally, and how it is this situatedness that shapes the interpretation, creation, and exchange of meaning. The consideration of the intricate entailments of this intercultural capability was extended by Steffensen et al. (2014) to include timescales and identity dynamics. The focus

specifically on identity formation was also taken up by Houghton (2013), with what she refers to as *savoir se transformer*.

In conceptualizing intercultural competence (or more precisely, "interlinguistic and intercultural practices and capabilities") for the purposes of assessment, Liddicoat and Scarino (2013) highlighted the need to capture:

- Observation, description, analysis, and interpretation of phenomena shared when communicating and interacting
- Active engagement with the interpretation of self (intraculturality) and "other" (interculturality) in diverse contexts of exchange
- Understanding the ways in which language and culture come into play in interpreting, creating, and exchanging meaning
- The recognition and integration into communication of an understanding of self (and others) as already situated in one's own language and culture when communicating with others
- Understanding that interpretation can occur only through the evolving frame of reference developed by each individual (pp. 130–131)

Assessment in this formulation, therefore while remaining focused on language and culture, encompasses more than language. It is at once experiential, analytic, and reflective. For Liddicoat and Scarino (2013), it includes (a) language use to communicate meanings in the context of complex linguistic and cultural diversity, with a consideration of both personal and interpersonal subjectivities, (b) analyses of what is at play in communication that is situated within particular social and realities and how language and culture come into play in the practice of meaning-making, and (c) reciprocal reflection and reflexivity in relation to self as intercultural communicator and learner.

In addition to extensive work on conceptualizing the assessment of intercultural practices and capabilities, practical work has been and continues to be undertaken to develop ways of eliciting these practices and capabilities (e.g., see, Byram 1997; Deardorff 2009; Lussier et al. 2007). Sercu (2004) attempted to develop a typology of assessment tasks including five task types: cognitive, cognitive-attitudinal, exploration, production of materials, and enactment tasks. This framework, however, does not address precisely these capture intercultural practices and capabilities.

As indicated, it is the alternative qualitative assessment paradigm, particularly within a hermeneutic perspective (Moss 2008) and inquiry approaches (Delandshere 2002), which offers the most fruitful basis for considering the assessment of these practices and capabilities in language/s learning. Liddicoat and Scarino (2013, chapter 8) discussed and illustrated ways of eliciting the meanings that learners make or accord to phenomena and experiences of language learning, and their analyses and reflections on meaning-making. The learner is positioned as performer and analyzer, as well as being reflective. An issue that remains to be considered with respect to elicitation is the complex one of integrating the performative, analytic, and reflective facets.

The area of judging is possibly the most complex of all, not only because educators hesitate to assess learner subjectivity and the realm of values and dispositions but also because of the difficulty of bringing together, in some way, the diverse facets of intercultural practices and capabilities. Although a framework for setting criteria for judging performance has been proposed (see Liddicoat and Scarino 2013, pp. 138–139), the extent to which criteria can be pre-specified or else should emerge from the specific context of the exchange still needs to be addressed.

Finally, there is not yet in the field a frame or frames of reference for making judgments of such practices and capabilities. The Council of Europe has sought to develop a scale to address this absence but efforts to date have not succeeded. This is not surprising given the complexity that this would entail. Although making judgments remains an area of uncertainty for assessors, it is not likely to be resolved by a generalizing scale.

## Multilingual Assessment Approaches

"Multilingual assessment" is a practice proposed by Shohamy (2011) that would take into account all the languages in the multilingual speaker's repertoire as well as "multilingual functioning" (Shohamy 2011, p. 418). Given the interrelationship between language/s and culture/s discussed above, this multilingual functioning also implies *multicultural* functioning. It is useful to distinguish at least two senses of multilingual assessment. The first is multilingual in the sense that multiple languages are available to the student, even though the assessment may be conducted in multiple but independent languages. The second is multilingual in the sense that student's performance reveals certain practices and capabilities that characterize the use of multiple languages by multilingual users as they negotiate, mediate, or facilitate communication. Although emanating from different contexts of language education and incorporating different terms, it is possible to draw some parallels between the more recent understandings of the assessment of intercultural practices and capabilities and the notion of multilingual functioning. Studies in assessment have been undertaken in relation to the first, but, although research on actual practices of multilingual speakers has been conducted, it has not been specifically in the context of assessment. Though not explicitly foregrounded, culture/s as well as language/s is at play.

Considering the first sense of "multilingual assessment," in an 8-year system-wide study in the multilingual context of Ethiopia, Heugh et al. (2012) demonstrated the value of learning and assessment in the student's mother tongue in bi-/trilingual teaching programs. Heugh et al. (2016) draw attention to bilingual and multilingual design of large-scale, system-wide assessments of student knowledge in two or three languages, as well as the unanticipated use, on the part of students, of their bilingual or multilingual repertoires in high-stake examinations. In the research reported by Shohamy (2011), immigrant students from the former USSR and Ethiopia, when assessed in Hebrew as the language of instruction in Israeli Jewish schools,

performed less successfully than the local, native students. Such students bring prior academic and cultural knowledge to the assessment situation, but this knowledge is not captured when the assessment is conducted in a language and culture that is different from that of their primary socialization. Furthermore, as Shohamy explained, these students naturally continue to use the linguistic and cultural resources developed prior to immigration, but their capacity to use this knowledge is not assessed. In these circumstances, the picture of their multilingual and multicultural achievements is distorted.

Cenoz and Gorter (2011) also highlighted approaches that draw on the whole linguistic repertoire of multilingual speakers. They reported on an exploratory study of students' trilingual written production in Basque, Spanish, and English in schools in the Basque Country. They focused specifically on the interaction among the three languages. The study showed that consideration of writing performance across three languages revealed similar patterns in writing skills in the three languages. They also illustrated that students use multilingual practices in creative ways and that achievement is improved when practices such as codeswitching and translanguaging are employed. These practices are linguistic and also cultural.

## Work in Progress

At the present stage of development, work in progress tends to be in individual, small-scale studies rather than part of large-scale programs of research and development. Conceptual work on modeling intercultural (or more precisely interlinguistic and intercultural) practices and capabilities will continue, as will consideration about the assessment of multiple languages and cultures and their relationship. Equally, discussion will continue about the assessment of capabilities beyond the linguistic (such as the capability to decenter or the capability to analyze critically or self-awareness about one's own linguistic and cultural profile). The Council of Europe's continuing work on the Common European Framework of Reference will seek to include indicators of intercultural competence because of the current desire to develop scaled, quantified levels of competence in all aspects of education. The current general education project of the Council of Europe, entitled "Competences for Democratic Culture and Intercultural Dialogue" (https://www.coe.int/t/dg4/education/descriptors_en.asp), may contribute to this line of development. Such quantification, however, runs counter to the qualitative, descriptive orientation that capturing these practices and capabilities entails.

An increasing range of research is being undertaken with a focus on multilingual functioning, especially processes such as translanguaging (Li Wei 2014; García and Li 2014). An explicit focus on the cultural and intercultural along with the linguistic and interlinguistic may add value to these research endeavors.

Some small-scale studies provide examples of work in progress. In a longitudinal study entitled "Developing English language and intercultural learning capabilities,"

Heugh (personal communication, October 2015) is incorporating translanguaging practices in the teaching, learning, and assessment of the English language of international students. The study involves practices in which students are invited to use their knowledge and expertise in their primary language in the process of developing high-level proficiency in English. Diagnostic assessment of students' written texts in Cantonese, Putonghua, and English allows for a more nuanced understanding of students' holistic capabilities in both their primary language and English (see Heugh et al. (2016)). This work is very much in line with Shohamy's (2011) desire that assessment recognizes the legitimate use and mixing of multiple languages, for it permits multilingual students to use their full linguistic, cultural, semiotic, and knowledge repertoires to interpret and create meaning. Heugh's work is demonstrating that these Chinese-speaking students also experience enhanced metalinguistic awareness of their own linguistic, cultural, and knowledge repertoire.

At the School of Oriental and African Studies, Pizziconi and Iwasaki (personal communication, October 2015) are researching the assessment of intercultural capabilities in the teaching and learning of Japanese. This work is being undertaken in the context of the AILA Research Network on Intercultural Mediation in Language and Culture Teaching and Learning. The project follows the development of linguistic and intercultural mediation capabilities in 14 learners of Japanese language before, during, and after a year of study in Japan. Through a variety of instruments, they are examining how students interpret, respond to, and negotiate identities, stereotypes, intercultural similarities and differences, the tensions arising from novel contact situations, the nature of the connections established, and how this is reflected in their language use. In short, they are investigating whether and how this long-term experience of "otherness" affects both performance and awareness.

Within the same network Angela Scarino, Anthony Liddicoat, and Michelle Kohler are developing specifications for the assessment of intercultural capabilities in languages learning in the K–12 setting in Australia. These will be used with teachers working in a range of languages to develop assessment procedures, implement them, and analyze samples of students' works for evidence of intercultural capabilities.

The new national curriculum for language learning in Australia has proposed an intercultural orientation to language teaching, learning, and assessment. Several studies related to the implementation of this curriculum, and related assessment practices, are currently being undertaken at the Research Centre for Languages and Cultures at the University of South Australia, in addition to experimenting with the design of elicitation processes.

The line of research by Cenoz and Gorter (2011) on trilingual students' participation in language practices that are shaped by the social and cultural context in the Basque Country and Friesland is continuing (see Gorter 2015) as is the work of Heugh et al. (2016).

## Problems and Difficulties

In expanding the construct of communicative competence toward symbolic, intercultural, and multilingual orientations (among the many new formulations that seek to represent this expansion), there is a need for explicit consideration of peoples' situatedness in the language/s and culture/s of their primary and ongoing socialization in the distinctive contexts of linguistic and cultural diversity. This attention is central to an understanding both of culture in language assessment and the role of culture in the assessment of students' learning outside the languages of their primary socialization, in multilingual and multicultural contexts. Difficulties remain at the level of conceptualization, elicitation, and judging.

## Conceptualizing Culture and Language Assessment

Further work is needed in conceptualizing the assessment of culture and the role of culture, particularly in multilingual and multicultural assessments. This may include, but is not limited to, the use of multiple languages in the assessment of content knowledge, the use of multiple languages and cultures in contemporary communication on the part of multilingual users, and a focus on interlinguistic and intercultural practices and capabilities in the assessment of additional languages. Both the conceptual work and its translation into assessment practice remain challenging because of the monolingual bias of both traditional SLA (May 2014; Leung and Scarino 2016) and traditional assessment (Shohamy 2011).

As part of this conceptual work, further consideration will need to be given to the context of culture and how it is perceived by participants in communication. Questions are being raised about the feasibility of assessing dimensions that go beyond the linguistic and the cultural, whether or not assessment philosophies and approaches can encompass the elicitation and judging of such complex practices and capabilities that go well beyond the linguistic and cultural per se, and the ethics of seeking to assess the realm of personal values, dispositions, effect, and critical awareness.

## Elicitation

The traditional product orientation of assessment does not capture the processual and reflective dimensions of assessing interlinguistic and intercultural and multilingual practices and capabilities. Finding productive ways of capturing cultural and intercultural interpretations will be difficult, and, in this regard, inquiry and hermeneutic approaches are likely to be of value (Moss 2008). These would permit the capturing for the purposes of assessment not only of experiences of interlinguistic and intercultural communication but also students' understandings of and reflections on the processes of meaning-making. The use of portfolios or journals, captured over time and including reflective commentaries, would seem fruitful. The complexity of

seeking to elicit the multiple facets of interlinguistic and intercultural communication (i.e., performance, analysis, and reflection) in an integrated and holistic way remains an area for experimentation. This is an important area for language educators who are concerned with developing as well as assessing such practices and capabilities. The elicitation process is necessarily framed by some understanding of the evidence that educators might expect to see in students' performances. As the kind of evidence of this kind of language-and-culture learning goes well beyond the accuracy, fluency, appropriateness, and complexity of language use, the very nature of this evidence will also require further consideration.

## Judging

As indicated earlier, there is a difficulty in judging, because of the uncertainty that arises for educators about judging student subjectivies and values. In the current state of play with assessment, what is absent is a larger frame of reference that educators need to bring to the processes of making judgments. Any instance of performance needs to be referenced against a map of other possible instances, but at this time, such a map is not available. As well, working with the notion of fixed rather than emerging criteria and scales adds complexity to the process. Educators desire certainty, when in fact there will necessarily be a great deal of uncertainty. This uncertainty relates to the absence of a shared frame of reference (such as one that they might have for a skill such as writing), but there are no firm guidelines as to what constitutes evidence. Furthermore, instances of communication of meaning across languages will be highly variable contextually, and yet it is precisely this linguistic and cultural variability and the linguistic and cultural situatedness of the participants that is being assessed in culture and language assessment.

In all three areas – conceptualizing, eliciting, and judging – the resilience of traditional practices is a major difficulty. In research, it is clear that both large-scale and smaller, grounded, ethnographic studies will be needed, focused on the assessment of interlinguistic and intercultural and multilingual practices and capabilities. It will be particularly fruitful for work in progress to be shared, compared, and theorized across research groups, given the immense diversity of local contexts of language-and-culture learning and its assessment.

Having highlighted the resilience of traditional assessment practices and their monolingual and monocultural bias, teacher education becomes a complex process of unlearning and learning. Teachers' assessment practices are heavily constrained by the requirements of the education systems in which they work. These requirements tend to be designed for accountability purposes more than for educational ones; therefore the environment is often not conducive to the kind of alternative practices that the assessment of these capabilities will require (see Scarino 2013 for a detailed discussion).

Finally, it must be recognized that this kind work in assessment, both in terms of practices and research, will be resource-intensive and raise issues of practicability. However, what is at stake in considering culture and assessment is the very nature of

language learning and its assessment and doing justice to capturing and giving value to the learning and achievements of students who are developing their multi-/interlinguistic and multi-/intercultural capabilities.

## Future Directions

What is needed is a program of research, undertaken in diverse contexts, that considers the meaning-making processes of students in their multi-/interlinguistic work and multi-/intercultural work. These are likely to include processes such as decentring and translanguaging, mediating understanding across multiple languages, and paying greater attention to the positioning of students. Evidence might include analyses of moment-to-moment actions/interactions/reactions, conversations, or introspective processes that probe students' meanings; surveys, interviews, and self-reports; and reflective summaries and commentaries on actions, and reactions. Also needed is a focus on identifying and naming or describing the distinctive capabilities that can be characterized as multi-/interlinguistic and multi-/intercultural. These are the unique capabilities that bi-/multilingual students display as they move across diverse linguistic and cultural worlds. They are likely to include not only knowledge and skill but also embodied experience and their consideration of language/s and culture/s within that experience. Here it would become necessary to understand not only students' ideas but also their life worlds, their linguistic and cultural situatedness, and their histories and values; to understand the way these form the interpretive resources that they bring to the reciprocal interpretation and creation of meaning; and to understand both themselves (intraculturally) and themselves in relation to others (interculturally).

## Cross-References


- ▶ Assessing English as a Lingua Franca
- ▶ Assessing Second/Additional Language of Diverse Populations
- ▶ Using Portfolios for Assessment/Alternative Assessment
- ▶ The Common European Framework of Reference (CEFR)


## Related Articles in the Encyclopedia of Language and Education

Anne-Brit Fenner: Cultural Awareness in the Foreign Language Classroom. In Volume: Language Awareness and Multilingualism

Francesca Helm: Critical Approaches to Online Intercultural Language Education. In Volume: Language, Education and Technology

Yiqi (April) Liu: Popular Culture and Teaching English to Speakers of Other Languages (TESOL). In Volume: Language, Education and Technology

Brenday O'Connor, Norma González: Language Education and Culture. In Volume: Language Policy and Political Issues in Education

Robert O'Dowd: Online Intercultural Pedagogy and Language Education. In Volume: Language, Education and Technology

## References


Atkinson, D. (1999). TESOL and culture. *TESOL Quarterly, 33*, 625–654. doi:10.2307/3587880.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox et al. (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa: Ottawa University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bennett, M. (1986). Towards ethnorelativism: A developmental model of intercultural sensitivity. In M. Paige (Ed.), *Cross-cultural orientation: New conceptualisation and applications* (pp. 27–70). New York: University Press of America.

Block, D. (2003). *The social turn in second language acquisition*. Washington, DC: Georgetown University Press.

Block, D. (2014). Moving beyond 'lingualism': Multilingual embodiment and multimodality in SLA. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and bilingual education* (pp. 54–77). New York/London: Routledge.

Byram, M. (1997). *Teaching and assessing intercultural communicative competence*. Clevedon: Multilingual Matters.

Byram, M. (2014). Twenty-five years on – From cultural studies to intercultural citizenship. *Language, Culture and Curriculum, 27*, 209–225. doi:10.1080/07908318.2014.974329.

Byram, M., & Zarate, G. (1994). *Définitions, objectifs et évaluation de la compétence socio-culturelle*. Strasbourg: Report for the Council of Europe.

Byrnes, H. (2010). Revisiting the role of culture in the foreign language curriculum. Perspectives column. *Modern Language Journal, 94*, 315–336. doi:10.1111/j.1540-4781.2010.01023.x.

Canale, M., & Swain, M. (1980). A domain description for core FSL: Communication skills. In Ontario Ministry of Education (Ed.), *The Ontario assessment instrument pool: French as a second language, junior and intermediate divisions* (pp. 27–39). Toronto: Ontario Ministry of Education.

Canale, M., & Swain, M. (1981). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Cenoz, J., & Gorter, D. (2011). Focus on multilingualism: A study of trilingual writing. *Modern Language Journal, 95*, 356–369. doi:10.1111/j.1540-4781.2011.01206.x.

Chalhoub–Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*, 369–383. doi:10.1191/0265532203lt264oa.

Deardorff, D. (2009). Implementing intercultural competence assessment. In D. Deardorff (Ed.), *The Sage handbook of intercultural competence* (pp. 477–491). Los Angeles: Sage.

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record, 104*, 1461–1484.

García, O., & Li, W. (2014). *Translanguaging: Language, bilingualism and education*. Basingstoke: Palgrave Macmillan.

Geertz, C. (2000). *Available light: Anthropological reflections on philosophical topics*. Princeton: University Press.

Gorter, D. (2015). Multilingual interaction and minority languages: Proficiency and language practices in education and society. *Language Teaching, 48*, 83–98. doi:10.1017/S0261444812000481.

Gumperz, J. J., & Levinson, S. C. (Eds.). (1996). *Rethinking linguistic relativity.* Cambridge: Cambridge University Press.

Gutierrez, K., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher, 32*(5), 19–25. doi:10.3102/0013189X032005019.

Halliday, M. A. K. (1993). Towards a language-based theory of learning. *Linguistics and Education, 4*, 93–116.

Halliday, M. A. K. (1999). The notion of 'context' in language education. In M. Ghadessy (Ed.), *Text and context in functional linguistics* (pp. 1–24). Philadelphia: Benjamins.

Heugh, K., Benson, C., Yohannes, M. A. G., & Bogale, B. (2012). Implications for multilingual education: Student achievement in different models of education in Ethiopia. In T. Skutnabb-Kangas & K. Heugh (Eds.), *Multilingual education and sustainable diversity work: From periphery to centre* (pp. 239–262). New York/London: Routledge.

Heugh, K., Prinsloo, C., Makgamantha, M. M., Diedericks, G., & Winnaar, L. (2016). Multilingualism(s) in system-wide assessment: A politics of innovation and resistance. Special issue on 'Multilingual Assessment'. *Language and Education.*

Heugh, K., Li, X., & Song, Y. (2016). Multilingualism and translanguaging in the teaching of and through English: Rethinking linguistic boundaries in an Australian university. In I. Walkinshaw, B. Fenton-Smith & P. Humphries (Eds.), *English medium instruction in higher education in the Asia-Pacific. From policy to pedagogy.* Dordrecht: Springer.

Houghton, S. A. (2013). Making intercultural communicative competence and identity-development visible for assessment purposes in foreign language education. *Language Learning Journal, 41*, 311–325. doi:10.1080/09571736.2013.836348.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal, 70*, 366–372.

Kramsch, C. (1993). *Context and culture in language teaching*. Oxford: Oxford University Press.

Kramsch, C. (2004). Language, thought and culture. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 235–261). Malden: Blackwell Publishing.

Kramsch, C. (2006). From communicative competence to symbolic competence. *Modern Language Journal, 90*, 249–252. doi:10.1111/j.1540-4781.2006.00395_3.x.

Kramsch, C. (2009). Discourse, the symbolic dimension of Intercultural Competence. In A. Hu & M. Byram (Eds.), *Interkulturelle Kompetenz und Fremdsprachliches Lernen. Modelle, empirie, evaluation* [Intercultural competence and foreign language learning: Models, empiricism, assessment] (pp. 107–122). Tűbingen: Gunter Narra Verlag.

Kramsch, C. (Ed.) (2014). Teaching foreign languages in the era of globalization. *The Modern Language Journal*, *98*(Special issue 1), 296.

Lee, C. D. (2008). The centrality of culture to the scientific study of learning and development: How an ecological framework in education research facilitates civic responsibility. *Educational Researcher, 37*, 267–279. doi:10.3102/0013189X08322683.

Leung, C., & Scarino, A. (2016). Reconceptualising the nature and outcomes in language/s education. *The Modern Language Journal, 100*(Suppl 2016). doi:10.111/modl.123000026-7902/16/81-95.

Liddicoat, A. J., & Scarino, A. (2013). *Intercultural language teaching and learning*. Malden: Wiley-Blackwell.

Lussier, D., Ivanus, D., Chavdarova-Kostova, S., Golubina, K., Skopinskaja, L., Wiesinger, S., & de la Maya Retamar, G. (2007). Guidelines for the assessment of intercultural communicative competence. In I. Lazar, M. Huber-Kriegler, D. Lussier, G. S. Matei, & C. Peck (Eds.), *Developing and assessing intercultural communicative competence: A guide for language teachers and teacher educators* (pp. 23–39). Strasbourg/Graz: European Centre for Modern Languages and Council of Europe Publishing. http://archive.ecml.at/mtp2/publications/B1_ICCinTE_E_internet.pdf. Accessed 15 Nov 2015.

May, S. (Ed.). (2014). *The multilingual turn: Implications for SLA, TESOL and bilingual education*. New York/London: Routledge.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden: Blackwell Publishing.

Moss, P. A. (2008). Sociocultural implications for assessment: Classroom assessment. In P. A. Moss, D. C. Pullin, P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity and opportunity to learn* (pp. 222–258). Cambridge: Cambridge University Press.

Risager, K. (2007). *Language and culture pedagogy: From a national to transnational paradigm*. Clevedon: Multilingual Matters.

Sapir, E. (1962). In D. Mandelbaum (Ed.), *Culture, language and personality: Selected essays*. Berkeley: University of California Press.

Scarino, A. (2010). Assessing intercultural capability in learning languages: A renewed understanding of language, culture, learning and the nature of assessment. *Modern Language Journal, 94*(2), 324–329. doi:10.1111/j.1540-4781.2010.01026.x.

Scarino, A. (2013). Language assessment literacy as self-awareness. Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(Special issue 3), 309–327. doi:10.1177/0265532213480128.

Sercu, L. (2004). Assessing intercultural competence: A framework for systematic test development in foreign language education and beyond. *Intercultural Education, 15*, 73–89. doi:10.1080/1467598042000190004.

Shohamy, E. (2011). Assessing multilingual competences: Adopting construct valid assessment policies. *Modern Language Journal*, *95*, 418–429. doi:10.1111/j.1540-4781.2011.01210.x.0026-7902/11/418–429.

Steffensen, S. V., Michiko, U., & Kramsch, C. (2014). The ecology of intercultural interaction: Timescales, temporal ranges and identity dynamics. *Language Sciences, 41A*, 41–59. doi:10.1016/j.langsci.2013.08.006.

Wei, L. (2014). Who's teaching whom? Co-learning in multilingual classrooms. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and bilingual education* (pp. 167–190). New York/London: Routledge.

Whorf, B. L. (1940/1956). In J. B. Carroll (Ed.), *Language, thought and reality: Selected writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

# Assessing Meaning

James Enos Purpura

**Abstract**

The quintessential quality of communicative success is the ability to effectively express, understand, dynamically co-construct, negotiate and repair variegated meanings in a wide range of language use contexts. It stands to reason then that meaning and meaning conveyance should play a central role in L2 assessment. Instead, since the 1980s, language testers have focused almost exclusively on functional proficiency (the conveyance of functional meaning – e.g., can-do statements), to the exclusion of the conveyance of propositional meanings or implied pragmatic meanings. While the ability to use language to get things done is important, excluding propositional content from the assessment process is like having language ability with nothing to say, and excluding pragmatic meanings guts the heart and soul out of communication.

In this chapter, I review how L2 testers have conceptualized "meaning" in models of L2 proficiency throughout the years. This logically leads to a discussion of the use of language to encode a range of meanings, deriving not only from an examinee's topical knowledge but also from an understanding of the contextual factors in language use situations. Throughout the discussion, I also highlight how the expression and comprehension of meaning have been operationalized in L2 assessments. Finally, I argue that despite the complexities of defining and operationalizing meaning in assessments, testers need to seriously think about what meanings they want to test and what meanings they are already assessing *implicitly.*

**Keywords**

Assessing meaning • Assessing pragmatics • Assessing content

J.E. Purpura (✉)
Teachers College, Columbia University, New York, NY, USA
e-mail: jp248@tc.columbia.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_1

## Contents



## Introduction

Nonnative speakers use second or foreign languages (L2) to get/give information at school, to create and maintain relationships online, to get a glimpse into other cultures, or, more subtly, to decipher intentions in political discourse. In other words, they use their L2 to express a wide range of meanings within social-interpersonal contexts (e.g., a friend recounting a subway story), social-transactional contexts (e.g., a client resolving a problem with a bill), academic contexts (e.g., a student writing a term paper), professional contexts (e.g., a scientist giving a talk), and literary or imaginative contexts (e.g., a poet writing/reciting a poem at a poetry slam). Since the ability to effectively express, understand, co-construct, negotiate, and repair meanings is the quintessential quality of communicative success, it stands to reason then that meaning and meaning conveyance should play a central role in L2 assessment (Purpura 2004).

In the L2 use domains mentioned above, language serves to generate messages that embody a variety of simultaneously occurring meanings. First and foremost, messages contained in utterances or texts encode *propositional* or *topical content*. Thus, the *propositional or topical meaning* of utterances or texts is said to convey factual information, ideas, events, beliefs, conjectures, desires, and feelings and is presumed to be context-free or decipherable apart from a communicative situation (Gibbs 1994). These propositional utterances are open to scrutiny in terms of their factual accuracy or their true-value[1]. Propositional meanings in the literature have also been referred to as the literal, semantic, sentential, compositional, grammatical, linguistic, inherent, conventional, or locutionary meaning of utterances and are generally considered a reflection of an individual's substantive, topical, or disciplinary, domain specific, subject matter, or content knowledge. They are fundamental to all language use.

The expression of propositions in messages is also used to assert a person's agency and express his intentionality in communicative interactions (e.g., to persuade) (Bloom and Tinkler 2001). By encoding intended meanings, these messages are used by interlocutors to *perform speech acts* or *communicative functions* with reference to some language use context. Thus, messages in utterances or texts *also* encode a user's *intended or functional meanings*. We can say then that the

[1]See Donald Davidson's essays for a fascinating discussion of truth and meaning.

propositional content of a message conveys more than what is said with words; it also communicates intended or functional meaning relevant to a language use context. Intended or functional meanings have been referred to in the literature as conveyed, interactional, illocutionary, or speaker's meaning. Unlike propositional meanings, functional meanings depend on the context of language use for successful interpretation. Similar to propositional meanings, however, functional meanings are fundamental to all language use as they represent an individual's *functional proficiency.*

Finally, while messages emerge from, depend on, and embody representations of an individual's internal mental content and serve as a reflection of personal agency and intentionality, they do not occur in isolation; they exist within a given sociocultural and interactional context and are thereby shaped by and interpreted within that context. Given that communication depends on the participants' shared presuppositions, experiences, and situational associations, much of what occurs in language use is unstated or implied. As a result, these same messages embody yet other layers of meaning, referred to as *implied* or *implicational pragmatic meanings*.

Implied pragmatic meanings emerge, for example, when someone is offered red wine and the acceptance response is: *Hey, I'm Italian*. Explicit in this response is the expression of propositional content – nationality. However, the response is also used *in this context* to communicate the respondent's functional meaning (i.e., my interlocutor made an offer; I'm *accepting*). Conjointly with the propositional and functional meanings, the response subtly encodes layers of other implied meanings including (1) *situational meanings* (i.e., the response reminds my interlocutor of my ethnicity and the role of red wine in my culture and presupposes my interlocutor will interpret my indirect response as an acceptance *in this situation*, even though not explicitly stated), (2) *sociolinguistic meanings* (i.e., the response conveys familiarity), (3) *sociocultural meanings* (i.e., the response presupposes what is common knowledge about Italians in *our culture*), and (4) *psychological meanings* (i.e., the response conveys playfulness). These implied pragmatic meanings have been referred to as sociopragmatic, figurative, extralinguistic, or implicational meanings.

Implied pragmatic meanings can also emerge as a simple function of word order. Consider the propositional, sociocultural, and psychological meanings associated with the utterance "My niece got married and had a baby" as opposed to "My niece had a baby and got married." Consider also how these meanings might vary across different social contexts.

In sum, language is efficiently designed to convey propositional meanings through topical content, together with functional meanings and layers of implied pragmatic meaning relevant to some language use context (Purpura 2016). The interaction among topical knowledge, language knowledge, and context and, I would add, the sociocognitive features of task engagement enable nuanced communication. And while these simultaneous encodings of meaning joyfully provide the basis for humor or poetry, they also increase the risk of communication breakdowns or the miscommunication of intent. They also present L2 learners with daunting challenges and heartwarming joys of learning to use an L2.

In L2 assessment, especially with nonreciprocal tasks, the propositional messages conveyed by an interlocutor, along with other pragmatic meanings, might be

considered a manifestation of a person's topical and language knowledge, her understanding of context, and her sociocognitive abilities. This is especially true if the propositions are true and faithful representations of the external world, if communication goals are met, and if the language output is grammatically precise and appropriate for the situation. With reciprocal tasks, however, these same messages serve only to initiate the establishment of joint understandings, followed by the co-construction of meanings relevant to the context. Communicative success thus is a *joint* product of the co-construction of variegated meanings. Finally, for communication to be successful, interlocutors need to express their own representations of mental content, reconstruct mental content representations of their interlocutors, and jointly co-construct these meaning representations synchronically and diachronically in verbal or nonverbal behavior. As Bates (1976 cited in Seliger 1985) stated:

> Meaning is a set of mental operations carried out by the speaker, which the speaker intends to create in the mind of the listener by using a given sentence. Whether or not the speaker actually succeeds is a separate issue. (p. 4)

Although the communication of meaning through propositional content and context plays a central role in L2 communicative success, L2 testers have devoted surprisingly little empirical attention to this topic. Instead, they continue to produce assessments, which, in my opinion, over-attribute value to the well-formedness of messages and to the completion of the functional acts, and they under-attribute importance to the conveyance of substantive, relevant, or original content, the development of topical progressions, and the conveyance of implied pragmatic meanings. This, by no means, is meant to diminish the significance of linguistic well-formedness in contexts where communicative precision is needed, or the importance of ascertaining L2 functional ability; it is simply a reminder that the primary aim of communication is the exchange of meanings in context. Thus, language, meaning through content, contextual considerations, and the sociocognitive considerations of task engagement should figure prominently in the design and validation of all L2 assessments.

In this chapter, I will review how testers have conceptualized "meaning" in models of L2 proficiency, describing the role that meaning conveyance through content and context has played in L2 assessment. I will argue for a reprioritization of meaningfulness over well-formedness in L2 test design since the exclusion of meaning from models of L2 ability likens to having language ability with nothing to say. Finally, I will highlight some of the problems and challenges in assessing meaning.

## Early Developments

Although some early language testers have purposefully disregarded the importance of meaning in models of L2 proficiency, others have clearly acknowledged the critical role it plays in communication and have addressed meaning and meaning conveyance in characterizations of L2 proficiency. This reflects the fact that people

use language in systematic ways to exchange messages on a variety of topics in a wide range of contexts, and in that way, they use language to get things done.

In 1961 Lado proposed a model of L2 proficiency based on a conceptualization of "language" as linguistic forms, occurring in some variational distribution, that are needed to convey linguistic, cultural, and individual meanings between individuals. *Linguistic meanings* referred to the denotative or the *semantic meaning* of "dictionaries and grammars" and were "interpretable without recourse to full cultural reference" (p. 3). Currently, linguistic meanings would be referred to as the literal, semantic, or propositional meaning of a form, utterance, or text. Linguistic meanings were said to reside in the use of phonology, sentence structure, and the lexicon and context limited to that contained within a sentence. *Cultural meanings* referred to concepts or notions that are culturally bound and only interpretable within a specific speech community or culture (e.g., *tapas*, *English breakfast*). Currently, these would be referred to as pragmatic meanings. Finally, *individual meanings* for Lado referenced words or concepts that lay outside the culture per se, indexing personal associations, such as when the word *dog* carries positive or negative connotations based on an individual's past experiences. With respect to these meanings, Lado argued that language is based initially on the linguistic meanings of structures and their combinations in an utterance, followed by other contextually bound meanings (p. 6). His schematization appears in Fig. 1.

Despite Lado's visionary depiction of "language" as a system of meaningful communication among individuals, he prioritized discrete linguistic elements (phonology, syntax, lexicon) of language use (reading, listening, speaking, writing) when it came to assessment design. He thus organized assessments around discrete forms, rather than around rich communicative situations in which layers of meanings could be elicited and measured. Consequently, assessing the extent to which messages are

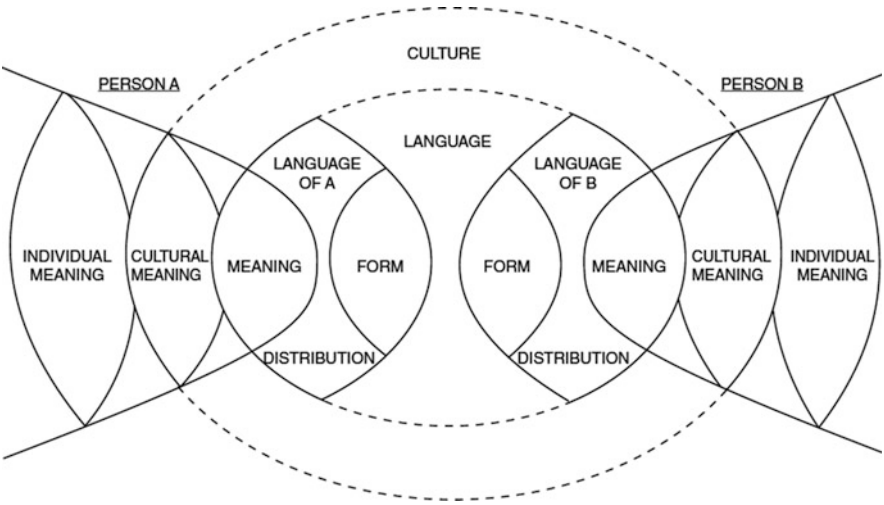


**Fig. 1** Language, culture, and the individual (Lado 1961, p. 6)

encoded semantically and communicated socially was secondary to the measurement of the linguistic resources used to deliver these messages. This perspective resulted in discrete-point assessments of forms and their associated meanings, instead of assessments eliciting propositionally accurate, topically elaborated, and situationally appropriate responses.

For example, in assessing phonological awareness through lexis, Lado presented students with two pictures, each representing a lexical item chosen because they happened to be minimal pairs (e.g., ship/sheep). Students then heard a word and selected the correct answer. In testing the meaning of counterfactual *if*-clauses, he presented examinees with a sentence and asked them to infer the correct propositional meaning of the sentence, based solely on the linguistic context, as seen below:

> If the windows were closed, I would ask you to open them.
> A) The windows are closed.
> B) The student goes to the windows and opens them.
> C) The student remains seated. (p. 158)

Given the minimal context, the inferencing needed to relate option (C) to the stem seems greater than the inferencing needed to understand the stem.

When it came to assessment, Lado generally preferred restricting test input to information that was "common knowledge in the culture where the language was spoken" (p. 205) and restricting the questions to selected-response items. This was based on the conviction that such restrictions would reduce the risk of introducing extraneous factors into the measurement process through situational context. However, when it came to extended production tasks, he argued that extraneous factors could be controlled to some extent by the use of rating scales revolving around linguistic difficulties and the success of meaning conveyance. Interestingly, the following language and meaning-based descriptors were used to rate the ability to narrate a story based on a picture:

> *2 – Conveys a simple description completely and correctly.*
> Conveys the simple description completely and correctly, but elaborates, and in so doing, makes some error, or error of vocabulary, grammar, or pronunciation – errors which interfere little with the understandability of the utterance. (Lado 1961, p. 240)
> *0 – Conveys very little meaning.*
> Conveys the wrong meaning.
> Makes errors, which obscure the meaning.
> Says nothing. (Lado 1961, p. 241)

While Lado is best known for the measurement of linguistic forms with discrete-point tasks, his conceptualization of L2 proficiency is clearly broader than that. From the onset, he recognized the importance of meaning in communication and provided recommendations for its measurement, not only in selected-response tasks, where meanings associated with grammatical forms and sentential propositions were assessed, but also in extended production tasks, where consideration was given to the overall conveyance of meaning in responses and to the extent that grammatical inaccuracy detracted from meaning conveyance. While much remained unspecified

in Lado's model regarding the types of meanings that were assessed or even the role of topical knowledge in enabling meaning conveyance, his ideas on this topic were insightful and, in my opinion, should have had a greater impact.

Carroll (1961/1972, 1968) also highlighted the role that meaning plays in language assessment. He defined "language" as:

> A system of 'rules' for generating utterances (or written representations thereof) that will be accepted by members of a given speech community as 'correct or 'grammatical' and understood by them as having a possible semantic interpretation. (Carroll 1968, p. 47)

Like Lado, Carroll recommended that L2 knowledge be specified in terms of linguistic forms, complemented by a semantic component. Unlike Lado, however, he recommended that less attention be paid to discrete morphosyntactic and lexical forms than to the "total effect of an utterance" or the "total meaning of the sentence" (p. 37). As a result, he proposed that that measurement of discrete components of L2 knowledge be supplemented by performance tasks that require the integration of components through connected discourse. Unfortunately, Carroll's inclusion of a meaning component in assessment was inconsistently applied in construct definition and operationalization.

Oller (1979) significantly advanced the conversation on "meaning" by describing "language" as both the interpretation and conveyance of factual content and the transmission of emotive or affective meanings in language use. He maintained "language is usually used to convey information about people, things, events, ideas, states of affairs, *and* [emphasis in the original] attitudes toward all the foregoing" (p. 17). He referred to the literal propositional meanings as the *factive information* of language use expressed by "words, phrases, clauses and discourse" (p. 33) and the psychological meanings of language use as *emotive* or *affective* information often carried by phonology or gestures. The emotive features were seen to "convey attitudes toward the asserted or implied state of affairs, [which] may further code information concerning the way the speaker thinks the listener should feel about those states of affairs" (p.18). Furthermore, according to Oller, the factive and emotive information of communication was highly dependent on the context of language use, which he referred to as (1) the *linguistic context*, consisting of the *verbal* and *gestural contexts* of language use, and (2) the *extralinguistic context*, involving the subjective and objective realities of "things, events, persons, ideas, relationships, feelings, perceptions memories, and so forth" (p. 19). He then asserted, "linguistic contexts are pragmatically mapped onto extralinguistic contexts, and vice versa" (p. 19). In other words, the meaning of the combined linguistic forms of an utterance (i.e., literal propositional meaning) is shaped by the pragmatic context in which the utterance or text is expressed. Also, the pragmatic mappings are a result of relating the propositional meanings of linguistic forms to extralinguistic context or human experience. In sum, Oller's farsighted conceptualization of L2 proficiency took into account linguistic knowledge, factual or topical knowledge, pragmatic knowledge (including emotive), and contextual features (extralinguistic) – a view strikingly similar to some current conceptualizations of L2 proficiency.

The following meaning recognition item illustrates how Oller attempted to measure the examinee's ability to decipher the meaning of the word *dropped* when the literal meaning was extended to suit the context:

John dropped the letter in the mailbox.
A) John sent the letter.
B) John opened the letter.
C) John lost the letter.
D) John destroyed the letter. (Oller 1979, p. 46)

In other words, "*dropping* a letter in a mailbox" is assumed to mean *sent*, not *let fall*, based on information in the context. Meaning extension here again derives from the available distractors, as an option such as "*put* the letter in the mailbox" would have been closer in meaning to the stem. Oller attempted to do the same in the following inferencing item.

| | |
|---|---|
| Man's voice: | Hello Mary. This is Mr. Smith at the office. Is Bill feeling any better today? |
| Woman's voice: | Oh yes, Mr. Smith. He's feeling much better now. But the doctor says he'll have to stay in bed until Monday. |
| Third voice: | Where is Bill now? |

A) At the office.
B) On his way to work.
C) Home in bed.
D) Away on vacation. (Oller 1979, p. 47)

Response (C) was also designed to measure the ability to decode meaning by mapping it onto an extralinguistic context (i.e., implied pragmatic meanings) as Bill's location cannot be derived *solely* from the linguistic context of the input, but from the presupposition that Bill's bed is in his home (i.e., he *could* have a bed in his office). Nonetheless, this item could have been a clearer example of meaning extension had distractor (C) been worded *At home*.

This approach to assessment supported Oller's proposal to use "integrative" or "pragmatic" tests to measure a learner's "pragmatic expectancy grammar," defined as a "psychologically real system that sequentially orders linguistic elements in time and in relation to the extralinguistic contexts in meaningful ways" (p. 34). An examinee would then display knowledge of pragmatic expectancy grammar by "relating these sequences of linguistic elements via pragmatic mappings to extralinguistic context" (p. 38). Importantly, pragmatic expectancy grammar aimed to connect the grammatical forms of an utterance, and the meaning expressed by this utterance in context, to some extralinguistic reality by inferential (i.e., cognitive) processes, thereby linking the utterance, I believe, to the individual's prior experience, knowledge, agency, and intentionality. Oller's position demonstrates a strong rejection of the then-current Bloomfieldian (1933) approach to linguistic analysis and formalism, reified by Chomsky (1957), where meaning was completely

disregarded from linguistic analysis,[2] in favor of a communication-based approach to language use.

In terms of measurement, Oller also recommended scoring protocols that specified not just "how well the text conforms to discrete points of morphology and syntax, but how well it expresses the author's intended meaning" (p. 386) in a given context, since judges always consider the communicative effectiveness of responses, whether or not they are scored.

Oller can be credited for highlighting not only the literal propositional (factual) and psychological (emotive/affective) content of utterances, encoded by linguistic forms, but also how these utterances relate to both internal mental states (i.e., cognition) and extralinguistic context. His work is also credited for specifying scoring methods that operationalize the assessment of propositional meaning conveyance in a variety of task types. Unfortunately, Oller never provided detailed theoretical or operational definitions of factive and emotive meanings conveyed in language use so that the quality of the factive information or the appropriateness of the emotive information in responses could be systematically assessed. Nor did he specify how test design could systematically account for extralinguistic context or the cognitive components of L2 proficiency in response elicitation. Nonetheless, Oller's insightful and forward-thinking ideas on meaning foreshadowed later conceptualizations of L2 proficiency.

Other testers have also highlighted the importance of meaning in language assessment. Inspired by Hymes (1967, 1972), Savignon (1972), Halliday (1973), Van Ek (1976), and Munby (1978), among others, Canale and Swain (1980) argued that language competence should be conceptualized within a framework of communication, where the functional meaning of utterances is central to L2 proficiency. In other words, priority was placed more on an individual's ability to achieve a communicative goal – to convey intended or functional meanings in context, than on the capacity to communicate accurate or relevant propositional content within the function. Secondary priority was given to an individual's ability to communicate with grammatical accuracy in ways that are socioculturally appropriate.

Canale and Swain' model conceptualized communicative competence as "a synthesis of knowledge of basic grammatical principles, knowledge of how the language is used in social contexts to perform communicative functions, and knowledge of how utterances and communicative functions can be combined according to the principles of discourse" (p. 20). They defined this construct in terms of grammatical, sociolinguistic, and strategic competence – later discourse competence was added (Canale 1983). While not the primary focus, the importance of meaning was noted in many parts of the model. For example, grammatical competence was defined in term of rules of semantics associated with "word meaning and sentence

[2]Surprisingly, the commitment to a syntactocentric approach to assessment, where only features of the language are assessed for accuracy, complexity, range, and fluency, has persisted in many assessments. As a result, the effective communication of propositions and the communicative meanings associated with these propositions are often ignored in the measurement process.

meaning" or the notion of "getting one's point across" (p. 10) (i.e., propositional meaning), and sociolinguistic competence was described as the sociocultural rules of language use and the rules of discourse (i.e., pragmatic meaning) (see Halliday and Hassan 1976; van Dijk 1977; Widdowson 1978). Canale and Swain further argued that learners need to know both sets of rules in order to appropriately express and understand meanings, especially when there is a "low level of transparency between the literal meaning of an utterance and the speakers' intended meaning" (p. 30) – in other words, in situations where the propositional content of utterances along with the communicative intents can be derived only from situational factors. Canale and Swain explained that the sociocultural rules of language use made possible the expression and interpretation of appropriate attitudes and registers within sociocultural contexts and that the discourse rules[3] allowed for the expression and interpretation of cohesion and coherence. Cohesive rules related forms to different types of referential meanings in texts,[4] while coherence rules related propositions and their communicative functions in sequenced discourse to implied rhetorical meanings in text. To exemplify, consider the implied rhetorical meanings created in following discourse sequence.

| Dialogue | Functional (and propositional) meanings | Implied rhetorical meanings (coherence) |
|---|---|---|
| A) That's the telephone | Device identification (*The phone is ringing*) | Implied request (*Can you answer the phone*?) |
| B) I'm in the bath | Expression of location (*I'm in the tub, presumably taking a bath*) | Implied refusal (*I'm taking a bath so I can't answer the phone*) |
| A) OK (Data from Widdowson 1978, p. 29) | Acknowledgment (*I acknowledge you are in the tub taking a bath*) | Implied acceptance of refusal (*I acknowledge you can't answer the phone*); implied response to request (*I'll answer it*) |

Canale and Swain's widely accepted model significantly broadened our understanding of the individual components of communicative competence and helped further the shift in assessment from a focus on grammatical forms to an emphasis on functional meanings in social interaction. It also highlighted, at least theoretically, the need to consider the sociolinguistic meanings carried by utterances, where an assessment might measure sociocultural appropriateness. Finally, it underscored the need to account for the rhetorical meanings encoded in cohesion and coherence. Although this model downplayed the role of topical knowledge and context in

[3]Canale (1983) later recognized that the rules of discourse might better be separated from the sociocultural rules of language use. Thus, he broadened the original conceptualization of communicative competence to include grammatical, sociolinguistic, and discourse competence and the cognitive component of language use, strategic competence.

[4]For example, anaphoric reference to relate the pronoun, *him*, to a referent, *boy*, or the logical connector, *then*, to relate temporality between clauses.

functional proficiency, it still inspired other testers to refine later notions of communicative competence as a basis for assessment.

## Major Contributions

Influenced by Canale and Swain (1980) and many others, Bachman (1990) proposed a model of communicative language ability framed within the notion of language use. This model was later refined in Bachman and Palmer (1996, 2010). In this model, meaning played a prominent role. Bachman and Palmer (2010) defined "language use":

> . . .as the creation and interpretation of intended meanings in discourse by an individual, or as the dynamic and interactive negotiation of meaning between two or more individuals in a particular situation. In using language to express, interpret, or negotiate intended meanings, language users create discourse. This discourse derives meaning not only from utterances or texts themselves, but, more importantly, from the ways in which utterances and texts relate to the characteristics of a particular language use situation. (p. 14)

While Canale and Swain limited their discussion to a language user's "communicative competence," defined in terms of language knowledge components and strategic competence, Bachman and Palmer (2010) significantly broadened the construct by arguing that in addition to *language knowledge*, language users in the act of communication need to engage their topical knowledge, affective schemata, and strategic competence when presented with some real-life or assessment task. They further argued that it is the interaction between an individual's language knowledge and these other factors that enable the user to create and understand meanings through discourse. While Bachman and Palmer never really provided an explicit definition of "meaning" in their model, they engaged in a compelling discussion of the knowledge components underpinning the creation and comprehension of meanings in discourse.

Bachman and Palmer's (2010) comprehensive description of language use consisted of language knowledge, topical knowledge, affective schemata, strategic competence, and other personal attributes; however, I will limit this discussion to an examination of language and topical knowledge given their role in the communication of meaning.

Bachman and Palmer conceptualized *language knowledge* as the interactions between organizational and pragmatic knowledge. *Organizational knowledge* was defined as (1) the knowledge that users need to produce or interpret spoken and written utterances to construct meaning – i.e., *grammatical knowledge*, or knowledge of vocabulary, syntax, phonology, graphology, and (2) the knowledge they need to organize these utterances into coherent spoken or written texts – i.e., *textual knowledge*, or knowledge of cohesion and rhetorical/conversational organization. Although they did not explicitly frame organizational knowledge in terms of forms and their associated meanings, they alluded to these two dimensions in discussing

scoring. For example, when a Spanish learner says **hers dogs* instead of *her dogs*, the incorrect utterance reveals his knowledge of *cohesive meaning* (correct reference to a female) and lack of knowledge of *cohesive form* (possessive adjectives do not agree with nouns in number in English). Therefore, in these cases, they recommended assigning one point to meaning and zero to form.

The second component of language knowledge in this model, *pragmatic knowledge*, was defined as the mental representations needed to "enable users to create or interpret discourse by relating utterances or sentences and texts to their meanings, to the intentions of language users, and to the relevant characteristics of the language use setting" (Bachman and Palmer 2010, p. 46). Pragmatic knowledge was further defined in terms of *functional* and *sociolinguistic knowledge*. Both components deal with meaning on some level.

Functional knowledge was said to "enable us to [express and] interpret relationships between utterances or sentences and texts and the intentions of language users" (Bachman and Palmer 2010, p. 46) in order to accomplish some communicative goal in context. Interestingly, this definition characterizes functional knowledge as a feature of the co-construction of communicative goal between two or more individuals, rather than as an attribute of a single user's communicative intentionality. As a result, a learner might be seen as demonstrating evidence of functional knowledge by responding to a friend's question *Can I give you some more wine*, with *Sure*, *pour away*, instead of *Yes*, *you are strong enough to lift the bottle*. In this context, the learner interpreted her friend's question as an indirect offer, rather than a query about ability, thereby achieving communicative success. If she had responded with a confirmation of ability, rather than an acceptance, this might have confused the interlocutor, resulting in communicative failure (i.e., lack of functional knowledge), unless, of course, she was intentionally being sarcastic. Thus, functional knowledge enables users to utilize context, as minimal as it is, to reassign meaning from a literal proposition (*can* = ability) to an intended meaning (*can* = request), or even to an implicated meaning (*can* = sarcasm) based on the communicative function of the utterance in discourse. Functional knowledge is thus seen as enabling users to *get things done through language* (van Dijk 1977).

Drawing on Halliday (1973) and Halliday and Hasan (1976), Bachman and Palmer identified four categories of functional knowledge that permit users to communicate joint intentions: *knowledge of ideational functions* (i.e., use of functions to relate ideas related of the real world – informing), *knowledge of manipulative functions* (i.e., use of functions to impact the world around us – requesting), *knowledge of heuristic functions* (i.e., use of functions to extend their knowledge of the world – problem-solving), and *knowledge of imaginative functions* (i.e., use of functions related to imagination or aesthetics – joking). In each case, a user would be judged on her ability to perform these functions.

Functional knowledge thus embodies the mental structures needed to communicate contextually relevant intentions between users with respect to the four communicative goals. It also enables users to get things done through language, thus explaining its operationalization in assessments as *can-do* statements. What remains unclear is the role that propositional content plays in expressing the four functions.

It would not be hard to imagine a situation in which learners can use the L2 accurately (grammatical knowledge) to summarize a story (functional knowledge), but the information in the summary (propositional knowledge through content) is inaccurate. In other words, it seems possible to demonstrate functional knowledge without displaying topical knowledge. Also unclear is the role that context plays in the expression or interpretation of functional knowledge. For instance, the interpretation of an indirect request (manipulative function) or a joke (imaginative function) depends on context for meaning conveyance, in addition to topic. So, given that meaning in these instances is derivable primarily, and sometimes uniquely, from features of context, is it possible to communicate functional knowledge without accurate or relevant topical content related to these contextual features? I would argue then that assessments based solely on functional proficiency provide only a partial estimate of a person's proficiency and one that can result in miscommunication.

Bachman and Palmer then defined *sociolinguistic knowledge* as the mental structures required to "enable us to create and understand language appropriate to a particular language use settings" (p. 47). Sociolinguistic knowledge targeted the user's capacity to use genres, dialects/varieties, registers, natural or idiomatic expressions, and cultural references or figures of speech appropriately in context. Thus, users able to use register appropriately and flexibly in formal contexts would be scored high for appropriate and wide knowledge of registers. The sociolinguistic component emphasized a user's "sensitivity" to register variations, natural or conventional expressions, and other linguistic features with relation to their appropriate use in context. Of note, however, is that this component is framed in terms of user sensitivity to these features, rather than in terms of the user's ability to recognize and transmit these implicit meanings in context.

Implicit in Bachman and Palmer's notion of sociolinguistic knowledge is first the inherent potential that users have for extending meaning beyond what is literally indexed in discourse. For example, the ability to use the expression *Your wish is my command* appropriately in context extends beyond an understanding of the literal propositional or functional meanings of the expression; it also presupposes an understanding of the context of language use as it relates to the transmission of sociocultural meaning (genie in a bottle) and sociolinguistic meaning – power (unequal), imposition (no limits), and distance (near). Although Bachman and Palmer did not frame sociolinguistic knowledge in terms of meaning, their model clearly highlighted the importance of sociolinguistic knowledge as a feature of language use and provided a basis for further work on the assessment of pragmatic ability.

One of the most interesting features of Bachman and Palmer's (2010) work in terms of meaning, however, was their discussion of topical knowledge as a consideration in assessment design and operationalization. While previous researchers vaguely referred to propositional content encoded in messages, Bachman and Palmer provided a compelling discussion of what topical knowledge refers to, how it interacts with other features of language use, and how it might be assessed.

They defined *topical knowledge* (also referred to in the literature as content knowledge, knowledge schemata, real-world knowledge, overall literal semantic

meaning, propositional content, or background knowledge) as knowledge structures in long-term memory (LTM) – unfortunately without further specification. They argued that topical knowledge is critical to language use because it provides the information needed to use language with reference to the real world and, I would add, with reference to an individual's internal world, as in creative expression. They stated that while topical knowledge is separate from language ability, it is still "involved in all language use" (p. 41) and is a factor in all test performance. They also maintained that since "it may not be possible to completely isolate language ability from topical knowledge in some test tasks, no matter how we score test takers' responses" (p. 325), testers should consider topical knowledge in assessment. Finally, they added that when an individual's topical knowledge interacts with the topical content in task completion, it impacts difficulty.

To disentangle the relationship between language ability and topical knowledge in test design, Bachman and Palmer offered three specification options:

1. Define the construct solely in terms of language ability.
2. Define language ability and topical knowledge as a single construct.
3. Define language ability and topical knowledge as separate constructs (p. 217).

Option 1 refers to assessment contexts making claims only about a component of L2 ability – e.g., knowledge of form. This might involve tasks focusing only on the measurement of form (with the topical meaning dimension being controlled) – e.g., when examinees are asked to choose among allophones (/t/, /d/, /id/) or among different verb forms (enjoy + *work*, *works*, *working*). In these cases, most testers would argue that topical knowledge is not part of the construct; thus, only one component of L2 knowledge (i.e., knowledge of form) would be scored. I would argue, however, that topical knowledge, in the form of metalinguistic knowledge, would be engaged – even if it is implicit knowledge. Option 2 refers to contexts making claims about L2 ability and topical knowledge as part of the same construct – e.g., when an international teaching assistant, *assumed to have the required topical knowledge for task completion*, must give a presentation in the L2. Only one score is taken and interpreted as the ability to use L2 and topical knowledge to teach. This option confounds language ability and topical knowledge, as scores could be affected by deficiencies in either. Finally, option 3 refers to contexts making claims about both L2 ability and topical knowledge as different constructs – e.g., in language for specific purposes (LSP) contexts, where examinees need to display their ability to use L2 ability to communicate disciplinary content – e.g., an analysis of food chains in an ecosystem. In this case, topical knowledge is conceptualized as drawing on explicit declarative memory or, I might add, a network of facts, concepts, principles, and rules in semantic memory that are assumed to be separate from language ability. Bachman and Palmer provided an example of a rubric [only partially presented] designed to measure topical knowledge in this context.

| Levels of knowledge/ mastery | Description |
|---|---|
| 4 complete | **Evidence of**: ***complete knowledge of relevant topical information*** |
| | *Range*: evidence of unlimited range of relevant topical information |
| | *Accuracy*: evidence of complete accuracy throughout range |
| 2 moderate | **Evidence of**: ***moderate knowledge of relevant topical information*** |
| | *Range*: medium |
| | *Accuracy*: moderate to good accuracy within range (Bachman and Palmer 2010, p. 352) |

While Bachman and Palmer's model greatly advanced our understanding of topical knowledge in L2 assessment, several questions remain. The first relates to the composition of knowledge structures related to topical knowledge in LTM. Are these knowledge structures limited to a semantic memory for factual or disciplinary knowledge (Dehn 2008), or might these assessments require examinees to draw on other memory sources in task completion? A second question concerns the relationship between topical and L2 knowledge. If topical knowledge is needed to generate and understand propositions encoded in language, is it actually possible to communicate without topical knowledge? And is it ever possible to assess L2 knowledge without some form of topical knowledge? Similarly, is it possible to have pragmatic ability without knowledge of the contextual situation (episodic memory) (Dehn 2008)? Finally, if communicative language ability *always* includes topical knowledge on some level, along with L2 knowledge, contextual understandings, and cognitive processing factors, then shouldn't these four features *always* be specified in assessment tasks involving communication? After all, each can potentially moderate L2 performance. In Fig. 2 I have attempted to schematize design considerations relating to context, topical content, language, and cognition/disposition as potential moderators of L2 proficiency in task engagement.

Building on Bachman and Palmer (1996) and Chapelle's (1998) interactionist approach to construct definition, Douglas (2000) reexamined the role of meaning by problematizing the relationship between background knowledge and L2 ability in the context of LSP assessment. He argued that in LSP contexts, an examinee's background knowledge was, in addition to L2 knowledge and strategic competence, a critical contributor to *specific purpose language ability* (SPLA). As a result, he defined background knowledge as part of the SPLA construct. Douglas defined *background knowledge* as "frames of reference based on past experience" (p. 35) within a discourse domain – a conceptualization reminiscent of what Baddeley et al. (2009) refer to as semantic declarative memory (i.e., factual knowledge) and episodic memory (i.e., experiential knowledge) associated with past contexts, events, and episodes related to LSP contexts. This insight, in my view, extends to all L2 assessments, as prior topical knowledge *on some level* is fundamental to meaning making. What, I believe, will fluctuate in L2 use (and in assessments) is the *type* of

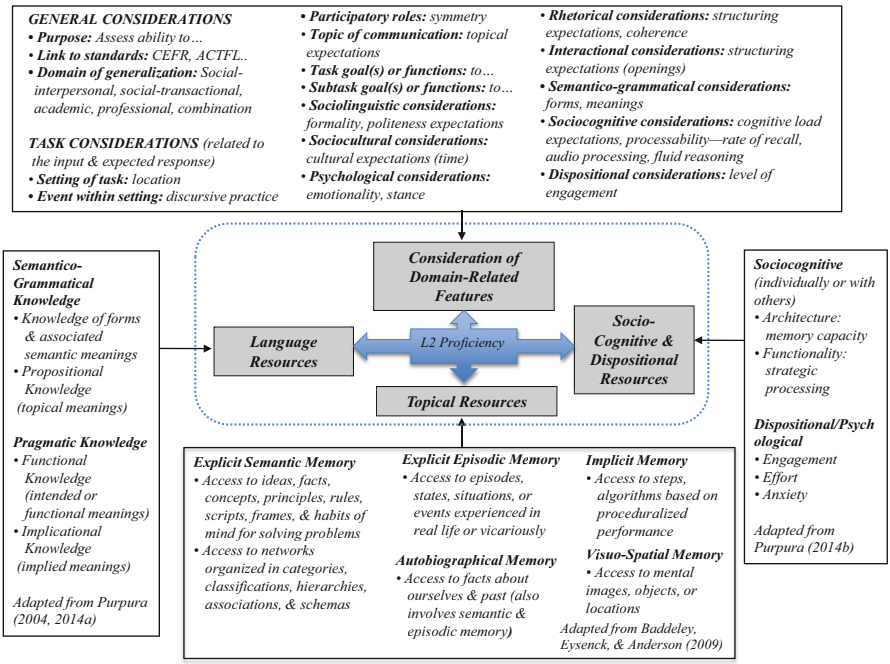


**Fig. 2** Representations of context, content, language, and cognition as potential moderators of L2 proficiency in task engagement

topical knowledge needed to communicate (e.g., disciplinary knowledge, autobiographical knowledge, situational knowledge), but certainly communication is impossible with no topical knowledge or flawed topical knowledge.

Finally, Douglas noted that even when LSP assessments are successful in engaging SPLA, performance on these tests is far too often based on scoring criteria revolving around the grammatical features of the output (e.g., intelligibility, fluency, grammatical accuracy, lexical appropriateness), rather than on "aspects of communicative language ability [. . .] deemed to be important" (p. 279) for inferences about SPLA. In other words, assessment criteria failed to target the examinees' ability to perform functions in LSP contexts that measure L2 ability in conjunction with critical aspects of disciplinary knowledge.

Building on prior research, Purpura (2004, 2014a, 2016) offered a slightly different conceptualization of L2 proficiency in which the ability to communicate meanings in some domain of L2 use depends upon the interaction between the context of language use, language knowledge, topical knowledge, and the sociocognitive and dispositional resources of task engagement, as seen in Fig. 2. In this conceptualization, meaning and meaning conveyance are seen as the cornerstone of L2 proficiency. This depiction of L2 proficiency is based on the assumption that proficiency, sampled as the simple utterance of a sentence with beginners to highly nuanced communication, requires a network of resources that enable users to

express, understand, dynamically co-construct, negotiate, and repair meanings, knowledge, and action, often in goal-oriented interaction. It also acknowledges the risks associated with meaning-related conversational breakdowns or flat-out miscommunications, due not only to semantico-grammatical deficiencies but also and more insidiously to pragmatic infelicities, which can easily lead to mutual misjudgment of intentions and abilities, miscommunication, and even cultural stereotyping (Gumperz 1999), which could ultimately promote linguistic manipulation, discrimination, and social inequity.

For example, considering a situation in which two L2 colleagues are preparing a presentation together in a café, successful communication would require (1) an understanding of the communicative goals and the sociocultural context of the meeting (situational understandings), (2) the use of semantico-grammatical resources (forms and semantic meanings), (3) the exchange of topical information (propositional meanings), (4) the accomplishment of interactional goals in talk-in-interaction (functional meanings), and (5) the nuanced communication of other implicated meanings relevant to the context (pragmatic inferences), such as a sense of camaraderie, collaboration, and comity. Finally, the ability to integrate these components in the goal achievement depends upon the users' (6) socio-cognitive mechanisms relating to the brain's architecture (e.g., memory), its functionality through processing (e.g., strategies), and (7) other dispositional factors (e.g., engagement, effort, attitude) (Purpura 2014b). In sum, successful communication in this context involves a complex network of interacting competencies, which can be assessed independently or as a whole, but each can potentially contribute to score variability.

In this *meaning-oriented model of L2 proficiency*, L2 knowledge depends on two mental assets: semantico-grammatical knowledge and pragmatic knowledge, both inextricably linked at the level of meaning in communication. *Semantico-grammatical knowledge* involves a user's knowledge of *grammatical forms* and their associated *semantic meanings* on the one hand and their ability to use these forms together to convey *literal propositional* or *topical meaning*. *Knowledge of grammatical forms* involves linguistic features at both the (sub)sentential (i.e., phonological/graphological, lexical, morphosyntactic forms) and the discourse levels (i.e., cohesive, information management, interactional forms). Knowledge of these forms has often been assessed in terms of *accuracy* or *precision*, *range*, or *complexity* or can also be inferred through characterizations of L2 production (i.e., percentage of error-free clauses) (see Ellis and Barkhuizen 2005).

*Semantic meaning* (also referred to as grammatical or literal meaning) is more complex. At the subsentential level, it encompasses the literal or propositional meaning(s) associated with individual forms. For example, semantic meaning on the subsentential level can be associated with the dictionary meaning of a lexical item, the morphosyntactic meaning of a past tense form (= past time, completed action), the referential meaning of a cohesive form (*hence* = conclusion), or the interactional meaning of a discourse marker (*anyway* = topic shift marker).

At the sentential level, however, grammatical forms along with their semantic meaning(s), arranged in syntax, conspire to produce the *literal propositional* or

*topical meaning* of the utterance. Literal propositional meaning encodes the topical content of a message and is often referred to as *factual*, *literal*, *topical*, *subject matter*, *domain specific*, or *disciplinary content*. Propositional meaning references subject matter literality, truth-conditional literality, or context-free literality (Gibbs 1994) and is available in LTM through topical knowledge by accessing (1) explicit semantic memory of facts, concepts, ideas, principles, rules, scripts, frames, or algorithms; (2) explicit episodic memory of states, episodes, situations, or experienced events; (3) autobiographic memory (Baddeley et al. 2009); and so forth (See Fig. 2). Some testers have vaguely referred to this as "general background knowledge." Interestingly, the literal propositional meaning of an utterance is its default meaning, especially when insufficient extralinguistic context is available for interpretation. Literal propositional meaning can be a source of ambiguity in indirect speech and is, amusingly, a critical part of puns (e.g., *A boiled egg in the morning is hard to beat*). With additional context, however, ambiguous propositional meanings often give way to the speaker's functional meaning in context for interpretation. Finally, the ability to convey propositional meaning depends on the user's ability to relate conceptual mappings available in LTM to situative contexts in order to generate propositional content (Pellegrino et al. 2001).

The propositional meaning of utterances or texts is often measured in terms of *meaningfulness* or *content control*, referring to the extent to which a user *gets her message across*, or the degree to which the topical content is accurate, relevant, sufficiently elaborated, and original. Propositional meaning can also be measured through *comprehension*, or the extent to which the *topical meaning of the message or text is understood*. Thus, the propositional or topical meaningfulness of utterances or texts encodes the user's expression or comprehension of content as it reflects a felicitous representation of the real world. Finally, in some assessment contexts, propositional meaningfulness is assessed via L2 production features such as the number of idea units encoded in texts (see Zaki and Ellis 1999).

The current scoring guide for the speaking section of the *TOEFL Primary* (ETS) provides a good example of how propositional knowledge has been operationalized in their scale descriptors.

| Language use, content, and delivery descriptors (TOEFL Primary) |
|---|
| **The test taker fully achieves the communicative goal** |
| A typical response at the 5 levels is characterized by the following |
| *The meaning is clear*. Grammar and word choice are effectively used. Minor errors do not affect task achievement. Coherence may be assisted by the use of connecting devices |
| *The response is full and complete. Events are described accurately and are easy to follow* |
| Speech is fluid with a fairly smooth, confident rate of delivery. It contains few errors in pronunciation and intonation. It requires little or no listener effort for comprehension (italics added) |

*Pragmatic knowledge* is the second component in this model and refers to knowledge structures that enable learners to utilize contextual factors such speech acts, indexicals, presuppositions, situational and cultural implicatures, and conversational and textual structuring to understand, express, co-construct, or negotiate

meanings *beyond what is explicitly stated* by the propositional meaning of the utterance. Pragmatic knowledge is multifaceted, but, for measurement purposes, can be defined in terms of the mental resources related to the communication of *functional* and *implied or implicated meanings* in language use. Thus, pragmatic knowledge depends on both a person's *functional knowledge* and her *implicational knowledge*. So, when a person wanting salt in a restaurant decides to formulate a message about this desire, her linguistic expression of it encodes the propositional meaning of the utterance. *Simultaneously*, her message in this context functions as a request, thereby encoding her agency and intentionality (Bloom and Tinkler 2001); it encodes functional meaning. The ability to understand and comprehend functional meanings in talk and text then depends on a person's *functional knowledge*, a critical component of pragmatic knowledge. Finally, as functional knowledge allows us to use messages to *get things done* in communication, this core competence has been operationalized to generate functional performance descriptors of L2 proficiency as seen in the *can-do* statements of the CEFR (Council of Europe 2001) (http://www.coe.int/en/web/portal/home), the TESOL Pre-K-12 Proficiency Standards Framework (TESOL 2006) (http://www.tesol.org/advance-the-field/standards/prek-12-english-language-proficiency-standards, the American Council on the Teaching of Foreign Languages guidelines (https://www.actfl.org/publications/guidelines-and-manuals/actfl-performance-descriptors-language-learners), and the Canadian Language Benchmarks (2012) (http://www.language.ca).

More interestingly, *pragmatic knowledge* also involves knowledge structures that enable learners to simultaneously encode, onto these same utterances or texts, a wide range of meanings that are implicated by shared presuppositions, experiences, and associations with reference to the communicative situation. This can be done through the select use of verbal and nonverbal resources in conjunction with a range of contextual factors. The ability to understand and comprehend these implied meanings in talk and text then depends on a person's *implicational knowledge*, another critical component of pragmatic knowledge. For example, the person in the restaurant, mentioned above, had a choice of making her request for salt in several ways. She could have been friendly, patient, and witty or aloof, demanding, and snide. These meanings can all be encoded in the simple request for salt. Given the complexities of pragmatic inference, these meanings often pose a serious challenge to L2 speakers and are clearly associated with L2 proficiency. In Fig. 3, I have identified the following seven types of implied pragmatic meanings encoded in talk and text (adapted from Purpura 2004, p. 91):

- ***Situational meanings***[5]: based on understandings of the local context of situation (i.e., how to communicate meanings *specific to a given situation*) – e.g., acceptable, appropriate, natural, and/or conventional use of indirect functions,

[5]In Purpura (2004) the term *contextual meanings* was used. The term *situational meaning* is now preferred as it attempts to codify meaning extensions derivable only from the local speech event (i.e., *you had to be there to get it*).

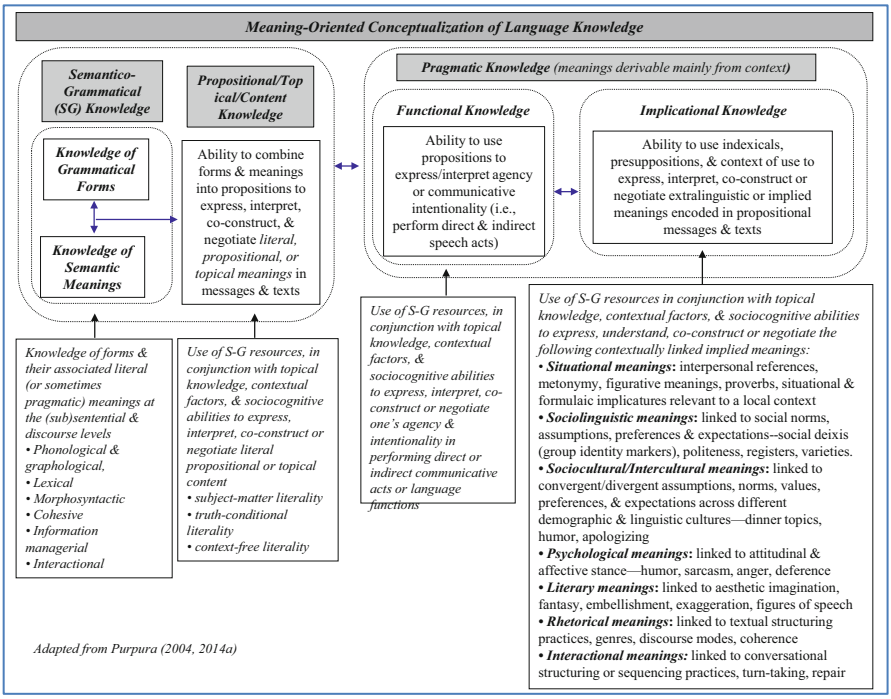


**Fig. 3** Meaning-oriented model of L2 knowledge (Adapted from Purpura 2004)

interpersonal references or associations, figures of speech, proverbs, and situational and formulaic implicatures

- ***Sociolinguistic meanings***: based on understandings of the social norms, assumptions, preferences, and expectations within a specific speech community (i.e., how to communicate with a given person *in a given social context*) – e.g., acceptable, appropriate, and conventional use of social deixis (group identity markers), politeness (relative power, degree of imposition, social distance), registers, varieties, etc.
- ***Sociocultural/intercultural meanings***[6]: based on understandings of the convergent or divergent assumptions, norms, values, preferences, and expectations *across* different demographic and linguistic cultures (how to communicate *within a given culture* or *across cultures*) – e.g., acceptable, appropriate, and conventional use of topic, humor, gratitude, regret, and criticism; avoidance of taboos; etc.

[6]Purpura (2004) specified only *sociocultural meanings*; however, as L2 communication in global contexts often involves speakers from diverse languages and cultures, the ability to understand and express *intercultural*, *cross-cultural*, or *transcultural meanings* was considered a pragmatic resource for intercultural communication.

- ***Psychological meanings***: based on understandings of affective stance (how to communicate mood, attitudes, feelings, emotionality, and other dispositions) – e.g., acceptable, appropriate, or conventional use of humor or sarcasm or the conveyance of anger, deference, patience, affection, self-importance, etc.
- ***Literary meanings***: based on understandings linked to aesthetic imagination, fantasy, embellishment, exaggeration, and figures of speech – e.g., appropriate, creative, and original use of literary conventions
- ***Rhetorical meanings***: based on understandings of textual structuring practices, genres, discourse modes, and coherence – e.g., acceptable, appropriate, and conventional use of organizational patterns
- ***Interactional meanings***: based on understandings of conversational structuring practices, sequencing practices, turn-taking practices, and repair practices – e.g., acceptable, appropriate, natural, and conventional practices associated with conversational norms, assumptions, and expectations

To summarize, pragmatic knowledge refers to the mental structures underlying the ability to communicate functional and other implicational meanings. The ability to utilize these structures in the task completion, however, is more complex, as it involves *pragmatic ability*, or the capacity to draw on semantico-grammatical resources to express or interpret propositional meanings, which, when used in situated interaction, carry contextually relevant layers of implicational meaning. Since pragmatic knowledge is a fundamental component of L2 knowledge, pragmatic ability is elicited in *all* contextualized language use *no matter the level of L2 proficiency*. The components of pragmatic knowledge can be assessed separately, or in combination, for situational, sociolinguistic, sociocultural/intercultural, psychological, rhetorical, or interactional appropriateness, acceptability, naturalness, or conventionality.

In most assessments, the contextual features needed for tasks to *systematically* elicit implicational pragmatic meanings are often insufficient. An exception to this is Grabowski's study (2009), which investigated examinees' ability to use three implicational meanings (sociolinguistic, sociocultural, psychological) in the context of a reciprocal role-play speaking task. The task prompt specified a particular social role for each interlocutor to assume (e.g., neighbor-neighbor), a communicative goal (e.g., get the neighbor to turn down the music at night), background information on the relationship between the speakers (e.g., persistent tensions over noise), culturally relevant information (e.g., values related to territorial rights, noise, and social harmony), and information relevant to the interlocutors' affective dispositions (e.g., frustrated). Thus, the sociolinguistic considerations of task design involved power distributions, social distance relationships, and absolute ranking of imposition; the sociocultural considerations addressed cultural norms, assumptions, and expectations of the situation in the local culture; and the psychological considerations involved a directive to assume a particular affective stance (e.g., frustration). The test taker responses were scored for grammatical accuracy, semantic (propositional) meaningfulness, and sociolinguistic, sociocultural, and psychological appropriateness based on an analytic rubric. The results showed that, in fact, highly

contextualized tasks could be used to systematically elicit propositional meanings alongside a range of implicated pragmatic meanings, which could be consistently scored and scaled across multiple proficiency levels.

The studies presented thus far have conceptualized meaning and meaning transmission mostly from a sociocognitive approach, describing mental representations of meaning (semantic, propositional, functional, and implicational) in the heads of interlocutors as they communicate, so that individual performance consistencies can be scored independently. However, according to proponents of the *socio-interactional approach* to construct definition (described in Purpura 2016), the sociocognitive approach fails to fully account for communicative success, since communicative success involves the *joint co-construction of relevant and appropriate meanings* that emerge from individuals interacting on a moment-by-moment basis to perform some goal-oriented activity (McNamara 1997; He and Young 1998). In the sociointeractional approach, the capacity to communicate meaning is not so much seen as ability within an individual than as the co-construction of meanings created between interlocutors in interaction. Evidence of this is seen, for example, when one interlocutor collaboratively finishes another's sentence or when interlocutors jointly contribute to the development of a topic when telling a story. While it is true that the creation of meanings in interaction is often a joint product of both interlocutors, it is also true that interlocutors avail themselves of individual resources in the co-construction of these meanings. If one interlocutor has fewer resources, the joint product is likely to suffer. Similarly, if sociocultural or intercultural norms of participation require an asymmetrical pattern of interaction (e.g., teacher-student), the joint co-construction of meanings is unlikely to emerge effectively, possibly affecting test performance. Thus, the sociointeractional approach might be better characterized as both a sociocultural *and* psychological phenomenon, where successful meaning conveyance in interaction is located *within* and *across* individuals *inside* sociocultural contexts.

These observations present testers with the conundrum of what to assess in interaction. Do we attempt to assess each interlocutor's capacity to express or comprehend meanings; do we assess the meaningful product of co-construction achieved by interlocutors; or do we assess both? While the idea of assessing only the joint co-construction of meanings is problematic in most assessment contexts, this approach has succeeded in highlighting the need to assess interactional practices related to turn-taking, conversational structure, and so forth. In the end, the ability to use these interactional practices appropriately (or not) in interaction encodes, as we have seen, a host of pragmatic meanings (e.g., the sociocultural meaning of interrupting inappropriately or the intercultural meaning associated with translanguaging).

Finally, a focus on meaning has been the cornerstone of a task-based language assessment (TBLA) approach to construct definition, where assessment revolves around the examinee's ability to use language meaningfully to accomplish tasks, designed as contextualized, real-world activities (e.g., give a presentation). According to Norris et al. (2017), these activities are also designed to require learners to draw on complex cognitive skills and domain-related knowledge, typically

aligned with a task-based language teaching (TBLT) pedagogical framework (Norris 2009). In TBLA, the competences needed to perform tasks are not drawn from a theoretical model of L2 proficiency, but rather are taken from the specific knowledge, skills, and abilities needed to accomplish the task at different performance levels, what Jacoby and McNamara (1999) call "indigenous assessment criteria."

Task accomplishment in TBLA has been assessed in many ways. Skehan (1998) and most other SLA researchers have evaluated the extent to which the language produced by examinees in task completion displays the linguistic features of complexity, accuracy, and fluency. This syntactocentric focus is, in my view, confusing given TBLA's focus on meaning in task accomplishment and would be more consistent with task-based pedagogy if this linguistic focus were complemented by a meaning focus involving an examination of the propositional features of L2 production together with judgments relating to the examinee's communicative functional ability through the successful exchange of meaningful, relevant, and original content. If more subtle characterizations of task completion were needed or if the results of these assessments were used for formative purposes, then TBLA rubrics would need to consider a pragmatic component. After all, we might have completed the task, but in the process offended our interlocutors.

Finally, an excellent example of a task-based approach to measuring functional communicative ability is seen in the English Language Section of Hong Kong's Curriculum and Development Institute, where assessment is organized around criteria related to the accomplishment of a sequence of goal-oriented tasks. These tasks required examinees to use language *meaningfully* to accomplish tasks they would likely perform in real life. Interestingly, the assessment explicitly specified general and task-specific assessment criteria related to the conveyance of meaning. Evidence of general content control was defined in terms of *topical relevance*, *propositional appropriateness*, *topical coverage*, and *ideational creativity*/*originality*, as seen in Fig. 4.

| **General and Task-specific Criteria for Assessing**<br>Task 1 – The Most Beautiful Cities in the World<br>Subtask 2: Writing back to your email pal (Writing) | |
|---|---|
| **General criteria for assessing writing** | **Task-specific criteria** |
| ***Content***—demonstrating<br>• relevance of ideas to the topic<br>• appropriateness of ideas<br>• substantive coverage<br>• creativity and originality of ideas | ***Content***<br>• writer starts by thanking email pal for information on Seattle and asking for missing details<br>• writer describes Hong Kong<br>• no irrelevant or inappropriate content<br>• substantive content |

**Fig. 4** Task-based assessment criteria for content (http://cd1.edb.hkedcity.net/cd/eng/TBA_Eng_Sec/web/seta.htm)

## Work-in-Progress

Several researchers are currently working on the role of meaning in language assessments. Bae et al. (2016) have just published an interesting, although somewhat controversial, paper on the role of content in writing assessment. Defining "content" as "ideas or meaning expressed in writing" and as "the extent to which those ideas are elaborated, developed, logical, consistent, interesting, and creative as well as relevant to the task requirements" (p. 6), they examined the extent to which the content ratings on an L2 writing assessment could be explained by vocabulary diversity, text length, coherence, originality, and grammar. Modeling the direct and indirect effects of these variables by means of structural equation modeling, they found that a substantial proportion of the variability associated with original, reflective, and interpretative content could be explained by the sum of these five elements. Thus, examinees displaying higher levels of content control produced more "original, reflective, and interpretive texts," thereby conveying greater levels of topical understanding. Bae et al. concluded that in summative assessment contexts, where practicality is always a concern, the assessment of the content *alone* provided an empirically sound, meaningful, and sufficient measure of writing ability.

Timpe Laughlin et al. (2015), interested in developing an interactive pragmatics learning tool for L2 learners of English in the workplace, provided a systematic and comprehensive review of the role of pragmatics as a component of L2 communicative language ability. This review offered a basis for rethinking the pragmatic competence construct. Influenced by a meaning-oriented approach to pragmatic competence, they proposed a model that addressed two fundamental features of communication: interactive construction and context. They then explicitly specified a meaning space in which two interlocutors in a given sociocultural and situational context can be assessed on their display of five distinct but interrelated dimensions of L2 knowledge. These include sociocultural knowledge, pragmatic-functional knowledge, grammatical knowledge, discourse knowledge, and strategic knowledge. Finally, they provided several interesting examples of task types that could be used in the measurement and ultimate development of pragmatic-functional awareness of L2 learners.

Finally, drawing on the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL™) project (Bennett 2010; Bennett and Gitomer 2009) at ETS and on Sabatini and O'Reilly's (2013) application of this work to reading literacy assessments, Sabatini et al. (2016) proposed a technique for organizing online assessments to measure the students' ability to *display and develop* language and topical knowledge while performing a tightly structured and topically coherent sequence of tasks designed to guide them through the resolution of a goal-oriented problem within a real-life scenario (Sabatini and O'Reilly 2013). These scenario-based assessments thus endeavor to measure the extent to which learners, with different levels of background knowledge, understand topical content in written (reading ability) and spoken text (listening ability), develop deep language and topical understandings with targeted assistance (the development of language and topical knowledge), and then use the newly acquired topical information to perform writing and speaking

tasks related to the scenario goal. This assessment is designed to reflect the multifaceted processes people use when working in a group to research and solve a complex problem. For example, a scenario might ask an examinee, along with his virtual group members, to enter a travel contest in which they have to submit a video-recorded pitch of two possible educational trips. To complete this task, examinees have to research websites and summarize the advantages and disadvantages of taking these excursions, learn about their misunderstandings of the texts, remediate these misunderstandings, synthesize the findings, prioritize the advantages over the disadvantages, and provide *meaningful and content-responsible* recommendations for the best trip. This assessment thus provides a perfect opportunity for assessing the display *and development* of L2 proficiency, topical knowledge, and reasoning skills in which contextual factors, L2 resources, topical resources, sociocognitive, and dispositional resources convene to play an explicit role in task achievement.

## Problems and Difficulties

While many testers have recognized the critical role of language in expressing meaning in assessments, only a few have endeavored to define the construct in ways that would allow it to be measured systematically and meaningfully. This comes as no surprise as researchers have had difficulty defining meaning and its relationship to L2 proficiency. After all, two broad fields of linguistics, semantics and pragmatics, have grappled with the meaning of meaning for centuries, with no one coherent model. The fundamental challenge with meaning, in my opinion, is that utterances expressed in context do not encode one meaning; they naturally encode several layers of meaning as we have seen. Nonetheless, we all seem to recognize successful communication when we see it.

To illustrate these complexities, consider, at the subsentential level, we can assess the meaning of a phonological form (e.g., rising intonation to encode curiosity) or the meaning of a morphosyntactic form (e.g., past conditional form to encode regret). At the sentential level, we can assess the meaning of a proposition – a statement that can be true or false. However, this becomes really interesting in L2 contexts when similar meanings across languages are not expressed in the same way. For example, *I dropped my pen* in English would be *My pen fell from me* in Spanish and *I let my pen drop* in French. Then, when these same messages are uttered in context with other interlocutors, the mutually conveyance of meanings becomes much more nuanced and complex. *What the speaker said* (propositional meaning) and *intended to communicate with the message* (intended or functional meaning) is overshadowed by *what was achieved by the message* (functional meaning) and *what was implied by it* (implicational meanings). At this point, meanings depend on pragmatic inferences based on contextual factors. On one level, meanings are contingent upon the mental common ground they have established regarding a set of propositions each speaker takes for granted in that context (Portner 2006) or a set of shared contextual associations. For example, a speaker might use a proposition to accomplish some action (e.g., invite), thereby encoding propositional and functional

meaning. Similarly, she might also use the proposition to communicate nuanced subtexts relating to the social or cultural context or to speaker's psychological state of mind. What is complex is that these meanings are simultaneously encoded in contextualized utterances or texts.

The challenge then for testers is what meanings to assess and how to assess them. The answer, of course, depends on the assessment purpose. However, as Grabowski taught us, we can rest assured that by specifying the appropriate amount of context in the input, it is indeed possible to assess only one or all these layers of meaning systematically and meaningfully. The critical takeaway, then, is for testers to think about what meanings they want to test and to score appropriately. Testers need also to be conscious of the meanings they are assessing *implicitly*.

## Future Directions

The field of L2 assessment has long engaged in debates about how to define L2 knowledge and what components, other than L2 knowledge, contribute to L2 proficiency. Over the years, testers have learned to acknowledge how tasks, similar to those examinees are likely to encounter in the real world, have served to engage examinees cognitively in L2 use. This led Chapelle (1998) and Chalhoub-Deville (2003) to conclude that in addition to trait considerations (L2 knowledge and strategic competence), L2 performance assessments needed to seriously consider context and interaction. This paper carries this a step further, arguing that topical knowledge expressed through meaning conveyance is equally important and should *always* be specified on some level. It also maintains that the complexity of the construct and the challenges in eliciting meanings systematically should be no excuse for ignoring one of the most fundamental features of communication and therefore of L2 proficiency. In the end, we need to think about meaning in ways that move beyond simple measures of vocabulary knowledge. L2 learners really need to know if *what* they said, *how accurately* they said it, and *what they accomplished* in saying it were effective or not. They also need to know if, in communicating it, they were contextually, socially, culturally, emotionally, and interactionally appropriate.

## Cross-References


- ▶ Assessing Students' Content Knowledge and Language Proficiency
- ▶ Cognitive Aspects of Language Assessment
- ▶ Critical Language Testing
- ▶ History of Language Testing
- ▶ Task and Performance-Based Assessment
- ▶ The Common European Framework of Reference (CEFR)

## Related Articles in the Encyclopedia of Language and Education

Lorena Llosa: Assessing Students' Content Knowledge and Language Proficiency. In Volume: Language Testing and Assessment

Frank Hardman: Guided Co-construction of Knowledge. In Volume: Discourse and Education.

Silvia Valencia Giraldo: Talk, Texts and Meaning-Making in Classroom Contexts. In Volume: Discourse and Education.

Nick C. Ellis: Implicit and Explicit Knowledge About Language. In Volume: Language Awareness and Multilingualism.

Anne-Brit Fenner: Cultural Awareness in the Foreign Language Classroom. In Volume: Language Awareness and Multilingualism.

Fredricka L. Stoller, Shannon Fitzsimmons-Doolan: Content-Based Instruction. In Volume: Second and Foreign Language Education.

Diana Boxer and Weihua Zhu: Discourse and Second Language Learning. In Volume: Discourse and Education.

## References

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language testing in practice*. Oxford: Oxford University Press.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York: Psychology Press.

Bae, J., Bentler, P. M., & Lee, Y.-S. (2016). On the role of content in writing assessment. *Language Assessment Quarterly, 13*(4), 1–27. doi:10.1080/15434303.2016.1246552.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70–91. doi:10.1080/15366367.2010.508686.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st century* (pp. 43–61). New York: Springer.

Bloom, L., & Tinkler, E. (2001). The intentionality model and language acquisition: Engagement, effort, and the essential tension in development. *Monographs of the Society for Research in Child Development, 66*(4), 1–91.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart & Winston.

Canadian Language Benchmarks. (2012). For English as a second language. http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333–342). Rowley: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Carroll, J. B. (1961/1972). Fundamental considerations in testing for English language proficiency of foreign students. Paper presented at the conference on Testing the English Proficiency of Foreign Students, Washington, DC, May 11–12, 1961. Reprinted in H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 313–321). New York: McGraw Hill.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46–69). London: Oxford University Press.
Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.
Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.
Chomsky, N. A. (1957). *Syntactic structures*. The Hague: Mouton.
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.
Dehn, M. J. (2008). *Working memory and academic learning*. Hoboken: Wiley.
Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, UK: Cambridge University Press.
Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.
Gibbs, R. J. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge: Cambridge University Press.
Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking* (Unpublished doctoral dissertation). Columbia University, New York.
Gumperz, J. J. (1999). On interactional sociolinguistic method. In S. Sarangi & C. Roberts (Eds.), *Talk, work, and institutional order* (pp. 453–471). Berlin: Mouton de Gruyter.
Halliday, M. A. K. (1973). *Explorations in the functions of language*. London: Edward Arnold.
Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. London: Edward Arnold.
He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing* (pp. 1–24). Philadelphia: John Benjamins.
Hymes, D. (1967). Models of the interaction of language and social setting. In J. Macnamara (Ed.), *Problems of bilingualism. Journal of Social issues, 23*, 8–28.
Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Middlesex: Penguin.
Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*, 213–241.
Lado, R. (1961). *Language testing*. London: Longman.
McNamara, T. F. (1997). Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446–466.
Munby, J. (1978). *Communicative syllabus design*. Cambridge, UK: Cambridge University Press.
Norris, J. M. (2009). Task-based teaching and testing. In C. Doughty & M. Long (Eds.), *The handbook of language teaching* (pp. 578–594). Malden: Blackwell.
Norris, J. M., David, J., & Timpe Laughlin, V. (2017). *A framework for designing second language educational experiences for adult learners*. New York: Routledge.
Oller, J. (1979). *Language tests in schools: A pragmatic approach*. London: Longman.
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
Portner, P. (2006). Meaning. In R. W. Fasold & J. Connor-Linton (Eds.), *An introduction to language and linguistics* (pp. 137–168). Cambridge: Cambridge University Press.
Purpura, J. E. (2004). *Assessing grammar.* Cambridge: Cambridge University Press.
Purpura, J. (2014a). Assessing grammar. In A. J. Kunnan (Ed.), *Companion to language assessment* (pp. 100–124). Oxford: Wiley. doi:10.1002/9781118411360.wbcla147.
Purpura, J. (2014b). Cognition and language assessment. In A. J. Kunnan (Ed.), *Companion to language assessment* (pp. 1452–1476). Oxford: Wiley.
Purpura, J. E. (2016). Second and foreign language assessment. *Modern Language Journal, 100*(Suppl), 190–208.
Sabatini, J., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling reading*

*comprehension: Behavioral, neurobiological, and genetic components* (pp. 100–111). Baltimore: Brookes Publishing.

Sabatini, J., O'Reilly, T., & Purpura, J. E. (2016). *Scenario-based language assessments in educational settings: Theoretical foundations, prototype examples and future applications*. Workshop presented at LTRC 2016, Palermo.

Sauvignon, S. J. (1972). *Communicative competence: An experiment in foreign-language teaching*. Philadelphia: Center for Curriculum Development.

Seliger, H. W. (1985). Testing authentic language: The problem of meaning. *Language Testing, 2*(1). doi: 10.1177/026553228500200102.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

TESOL (2006). TESOL Pre-K-12 Proficiency Standards Framework (http://www.tesol.org/advance-the-field/standards/prek-12-english-language-proficiency-standards

Timpe Laughlin, V., Wain, J., & Schmidgall, J. (2015). *Defining and operationalizing the construct of pragmatic competence: Review and recommendations (ETS Research Report No. RR-15-06)*. Princeton: Educational Testing Service. doi:10.1002/ets2.12053.

Van Dijk, T. A. (1977). *Text and context: Explorations in the semantics and pragmatics of discourse*. London: Longman.

Van Ek, J. A. (1976). *Significance of the threshold level in the early teaching of modern languages*. Strasbourg: Council of Europe.

Widdowson, H. G. (1978). *Teaching language as communication*. London: Oxford University Press.

Zaki, H., & Ellis, R. (1999). Learning vocabulary through interacting with a written text. In R. Ellis (Ed.), *Learning a second language through interaction*. Amsterdam: John Benjamins.

# Language Assessment in the US Government

Rachel L. Brooks

**Abstract**

The US government is one of the first and most influential language assessment organizations in the USA. With its foundation being the Interagency Language Roundtable (ILR) Skill Level Descriptions, the US government has developed and administered tests not only in proficiency skills (listening, reading, speaking, writing) but led the way in performance testing (translation, audio translation, and interpretation) and intercultural competence. The scope of testing in the US government is tens of thousands of tests administered annually in hundreds of languages. Important to the US government is its operational underpinnings; tests are developed and administered to meet the missions of the agencies. US government agency scores are used to make a wide range of high-stakes decisions that can impact not only the careers of the examinees but also the lives of people the world over. Tight deadlines and limited resources, as well as changing needs and complexities in language challenge government test developers. Research regarding US government language-testing examines issues such as the relationship between reading, writing and translation, rater characteristics, standard setting, and other topics meant to improve the quality of language testing. In recent years, the US government assessment programs have increased collaboration among agencies leading to additional resources and helping each agency better fulfill its mission.

**Keywords**

Government • Interagency Language Roundtable Skill Level Descriptions • Proficiency • Performance

R.L. Brooks (✉)
Federal Bureau of Investigation, Washington, DC, USA
e-mail: rachel.brooks@ic.fbi.gov.

© Springer International Publishing AG (outside the USA) 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_4

## Contents



## Introduction

Government testing programs span different types of agencies such as diplomatic, military, clandestine, and investigative. These agencies are responsible for administering their own language-testing programs, but they share resources and information, often under the umbrella of the Interagency Language Roundtable (ILR). The ILR provides a venue for agencies to exchange ideas, hold symposia, and share research (Jones and Spolsky 1975; ILR 2016). The US government collectively conducts tens of thousands of tests annually in nearly 200 languages, covering all levels of proficiency. The government conducts tests in a range of skills: listening, reading, speaking, writing, translation (including document, audio, and summary), interpretation, and transcription.

US government language testing poses unique challenges. Testing is tailored to operational needs that shift based on world events, impacting the types of tests needed and requiring tight deadlines. US government language testing is high stakes because it determines whether government personnel have a reliable ability to perform the language tasks to support defense, diplomatic, national security, and law enforcement needs. Testing programs meet these challenges by developing new tests, as well as adapting and adopting available resources for assessments. Testing not only impacts examinees but also the agency mission and, consequently, the citizens the agencies serve.

## Early Developments

In the US government, language learning and assessment programs have always focused on practical needs stemming from current events, such as wars, terrorist acts, and international events. Prior to the 1940s, the focus of language assessment was classroom assessments of reading proficiency. It was localized in each agency, with little interagency collaboration. The US involvement in World War II caused language training and testing efforts to increase significantly, leading to resource

sharing among agencies. Moreover, World War II shifted the focus of language learning from reading to listening and speaking. Radio transmissions became an integral part of wartime communication, leading to the need for foreign language intercepts. More and more soldiers were being deployed overseas, requiring conversational abilities. To meet these changes, Kaulfers (1944) outlined a methodology for aural and oral language evaluation, including rubrics and rating criteria. In 1949, the US Army released the first standardized tests of proficiency in reading, listening, writing, and grammar in 25 languages called the Army Language Tests (Pulliam and Ich 1968) based on Kaulfer's methodology.

The standardization of language testing also had an impact on language aptitude testing. Before World War II, US military language course placement was determined by a combination of measures, including IQ tests, general language aptitude tests, and tests of how well a person could speak a "first" language (Myron 1944). These tests were found to be ineffective measures of language aptitude once language training moved away from the translation method, leading to a formalized aptitude assessment (Petersen and Al-Haik 1976). One of two early aptitude tests was the Department of Defense's Defense Language Aptitude Test (DLAT). The Modern Language Aptitude Test (MLAT) followed the DLAT in 1959 and was widely used by agencies in both the USA and Canada. In 1976, the DLAT was revised, validated, and renamed the Defense Language Aptitude Battery (DLAB) (Petersen and Al-Haik 1976).

Before long, the Army Language Tests released in 1949 needed updating and in 1954 the Army Language School (now the Defense Language Institute Foreign Language Center (DLIFLC)) constructed the Defense Language Proficiency Tests (Pulliam and Ich 1968). Meanwhile, in 1952, the US Civil Service Commission was tasked with inventorying the language abilities of government employees across agencies, requiring standardized assessment criteria. Government personnel included native speakers, heritage speakers, and language learners, so a way to assess language proficiency regardless of how the language ability was attained was critical. The US government developed its own standardized criteria since no such criteria were found in academia (Herzog 2003; Jones and Spolsky 1975; Lowe 1985). The US Foreign Service Institute (FSI) of the Department of State came up with the first rating scale of functional language ability, with score levels 1–6. An independent testing office at FSI, established in 1958, extrapolated a format for reliable speaking testing from these criteria known as the "FSI test." In 1968, other US government agencies collaborated with FSI to develop and expand the criteria to cover speaking, listening, reading, and writing. This project resulted in the Interagency Language Roundtable (ILR) Skill Level Descriptions. Subsequently, federal government agencies worked to update and develop additional language tests based on the ILR. In particular, the FSI test was adapted for general proficiency use, expanding its breadth from the original FSI-focused scope, by a number of agencies and became known as the Oral Proficiency Interview (OPI) (Lowe 1988).

As the ILR Skill Level Descriptions were more broadly implemented across agencies, they received feedback and underwent revisions. The ILR scale adopted "plus" levels, which indicated language users with an ability that substantially

exceeded the base level, yet did not fully meet the next higher level. In 1985, the US Office of Personnel Management approved the ILR Skill Level Descriptions as the official criteria for evaluating the language proficiency of government personnel (Interagency Language Roundtable 1985). In the early twenty-first century, the ILR addressed the need to measure language in performance skills derived from operational language tasks such as translation, interpretation, transcription, and audio monitoring. The Translation and Interpretation Committee of the ILR joined with the Testing Committee to develop a set of performance skill level descriptions, including translation (2006), interpretation (2007), and audio translation (2011) (Brau 2013). Around the same time, discussions commenced on the importance of measuring the cultural knowledge and abilities used in communication between government personnel and native speakers overseas. To capture the progression of extralinguistic communication elements, the ILR developed the Skill Level Descriptions for Competence in Intercultural Communication (2012) (Interagency Language Roundtable 2016).

## Major Contributions

### Government Testing Criteria

The US government most often uses the ILR Skill Level Descriptions as their criteria for assessing language. The descriptions provide a common reference enabling organizations to have comparable expectations about general ability. They are an ordinal scale composed of six base levels from 0 to 5 with five plus levels from 0+ to 4+, totaling eleven ranges. They were developed by subject matter experts in language acquisition with experience in assessment representing the agencies that most frequently administer language testing (Lowe 1998). The ILR levels assume importance because most US government language tests use these scales as a reference. Therefore, they must be understood by all government stakeholders, including examinees, managers, training coordinators, etc. The descriptions do not provide comprehensive lists of abilities or linguistic functions and as such are subject to interpretation. The challenge in the production and use of the ILRs is that they must be general enough to meet the diverse needs of the agencies that use them, while being specific enough to control for reliable interpretation by the different organizations. The ILRs must meet the needs of the agencies that rely on them, which generally result in a lengthy development and approval process. Since the ILRs became the official language rating criteria for the US government, significant resources have been invested to develop and validate assessments based on them, including the Defense Language Proficiency Test (DLPT), the Oral Proficiency Interview (OPI), and the Verbatim Translation Exam (VTE). ILR-based tests look at a person's functional ability to perform linguistic job tasks specific to each agency and its validity lies in its ability to measure functional ability reliably. Agencies regularly conduct reliability checks from independent raters and have over the years

proved that the functional progression shown in the scales is accurate regardless of how the language was acquired (Brau 2013; Lowe 1988).

The ILR Skill Level Descriptions have importance outside the government context as well. They are the basis for the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines, which were intentionally designed to be commensurate and derivative of the ILR. As such, the ACTFL Guidelines are at times used within the US government context, such as in the Peace Corps and the Department of Education. Additionally, the ILR Skill Level Descriptions heavily influenced the NATO STANAG (standardization agreement) 6001 language proficiency guidelines, which are used by foreign governments, including Canada and several European countries (Bureau for International Language Co-ordination 2016).

The framework of the ILR Skill Level Descriptions has important ramifications for developing and scoring language proficiency tests. First, the ILR Skill Level Descriptions are non-compensatory, that is, strength in one feature cannot compensate for weakness in another feature at a given level. For example, someone who can orally support opinions on societal-level topics using precise vocabulary (a level 3 skill) cannot be considered to have an overall level of 3 in speaking if there are persistent errors that interfere with comprehension, such as failure to distinguish singular and plural. Second, overall control of functions, a person's ability to accomplish particular language tasks, rather than total absence of errors or perfection of understanding are important (Brooks 2013).

## Government Perspective

Since the major driving force behind government language testing is operational need, performance testing is essential. Within government contexts, the distinction between proficiency and performance testing has become significant. Proficiency testing refers to a holistic evaluation of a person's functional ability in the language. It is a general assessment that does not pay regard to how a language was acquired. The ILR scales for proficiency are the original four skills of listening, reading, speaking, and writing. When these first skill level descriptions were developed, testing focused on post language training exams. Assessing functional proficiency remains important because the government needs language generalists who have flexible language ability that can quickly meet needs. Government organizations highly value personnel who maintain high levels of general proficiency in a variety of skills.

In more recent years, it has become evident that testing of performance skills that require prerequisite proficiencies (i.e., translation which requires reading and writing proficiencies) is more practical than testing proficiency alone for government purposes. Performance tests, which measure a person's ability to perform a certain job, assess specific skills, such as translation, summarization, interpretation, and transcription all arise from operational tasks (Brau 2013; Child et al. 1991). Therefore, performance tests are a more practical and valid measure of the skills being used on

the job. Some agencies have worked to create performance tests since the late 1990s, but they are still only available in the top 30 or 40 tested languages. When performance tests are not available, testing programs have to rely on proficiency exams.

## Impact of Agency Mission

The US government agency that is probably most well known for foreign language training and testing is the Department of State (DOS), which includes the Foreign Service Institute (FSI). The School of Language Studies at FSI is responsible for foreign language training of foreign service officers who interact with counterparts in US embassies. Its personnel have regular contact with counterparts from numerous international backgrounds, requiring high-level language skills, particularly in speaking. Diplomats need to converse with foreign counterparts, read foreign documents, and listen to broadcasts in other languages. Language Services at the Department of State has translators and interpreters that routinely perform specialized language tasks such as translation of international treaties and agreements and interpretation of negotiations and official addresses. Translators and interpreters are expected able to understand nuance, tone, implied meanings, and cultural references. Moreover, employees of diplomatic agencies serve as the face of their country in foreign lands; therefore, miscommunication could potentially lead to serious ramifications on international relations. Consequentially, diplomatic personnel typically endeavor to communicate effectively and appropriately as educated native speakers of the foreign language. Skills such as negotiation, persuasion, tact, and other influencing skills are expected to be mastered. Language testing emphasizes speaking but also reading and listening for officers and translation and interpretation for linguists at Language Services. The testing program is geared to high-level proficiency, ILR levels 3 and above as a goal.

Within the Department of Defense (DOD), foreign area officers, like diplomats, work in embassies and may need to negotiate and communicate agreements in security cooperation efforts between the USA and other countries. Primarily, however, defense organizations focus on giving military personnel the communicative skills they need to survive in foreign lands. They teach speaking and listening in routine or survival communications, such as gathering information from residents about local activities and performing security operations. Other personnel may monitor recorded or written communications from hostile groups. Although military personnel often do not need high levels of proficiency, the stakes are high. Inaccurate transfer of information could lead to loss of life or property. The majority of those trained and tested at the DOD take listening, reading, and speaking proficiency tests at ILR levels 3 and below.

In clandestine services, such as the Central Intelligence Agency (CIA) and National Security Agency (NSA), agents working undercover need to develop structural competence, vocabulary, and pronunciation that are parallel to those of native speakers. Additionally, they must acquire native speakers' cultural and

pragmatic skills, so as to be indistinguishable from them. Language errors have the potential to lead to loss of life or intelligence. Agents gather intelligence through audio intercepts, so listening skills are paramount. Listening comprehension tasks are complicated by the inability to ask for clarification and by poor recording quality. Additionally, a large number of language tasks require decoding vague, accented, slang, and veiled language. Language testers work to interpret how this type of task fits into the general rating scales and how to reliably assess listening in such contexts.

Investigative and law enforcement agencies, such as the Federal Bureau of Investigation (FBI) and the Drug Enforcement Agency (DEA), generally, serve both criminal and intelligence missions. Operational requirements demand that language personnel have both monitoring and translation abilities, with added legal requirements that govern the collection of and reporting on evidence and intelligence. Monitors overhear and then write analytical summaries of information relevant to investigations, which are often distinct from the main idea or supporting details of the audio. National privacy laws restrict material that can be monitored, so audio is truncated, causing additional listening challenges. Documents that are collected as evidence for investigations need to be translated so that the information is accessible to agents working on the related cases. Translation errors can lead to the dismissal of evidence admitted in court proceedings. As in government organizations, most interpretation assignments are informal and involve interviewing speakers of other languages. Investigative agencies also employ undercover agents who are high-level speakers of foreign languages. In all of these cases, single skill testing does not sufficiently measure language for the task, therefore performance testing of combined skills is increasing. Inaccuracies in court interpretations can result in unwarranted imprisonment or unprosecuted crimes. High levels of proficiency in speaking and listening do not necessarily result in high-quality interpretation. Therefore, most court systems test for interpretation skills directly rather than inferring them from the results of speaking proficiency tests.

In the USA, the Department of Education (DOE) oversees school curricula, initiatives, and assessments in all subject matters, including language. Educational institutions use language testing and their corresponding frameworks to measure the progress of student language learning. Education personnel referring to rating scales are generally interested in the lowest levels offered, as the majority of students will achieve results at these levels. Combined skills such as interpretation and translation are not taught except in specialized schools; therefore, educational agencies refer largely to the scales for the four primary skills using the ACTFL Guidelines. Often outcomes on these tests are used to measure student achievement and teacher performance.

In the US Peace Corps, humanitarian volunteers serve for one or two years in foreign countries teaching language or providing aid services. Most language learning that is done is in country and addresses survival needs rather than professional contexts; therefore, participants typically only achieve low levels of language proficiency. As in educational departments, service personnel may be tested via speaking proficiency tests to measure how much language learning was achieved. In other cases, such as the US National Language Service Corps, volunteers are

reserves. They are tested for general speaking proficiency so that, when a need arises, the organization knows which volunteers are most capable.

Increasingly, almost all aspects of government work are affected by foreign languages and all government agencies need some types of language users. Border officers need to conduct basic interviews, but they also need to be able to detect if a person is being dishonest. The Internal Revenue Service investigates and audits tax records and payments, requiring language personnel with reading skills to review records kept in foreign languages and writing skills to issue official letters in a language that the recipient can understand. Census workers conduct surveys in multiple languages to ensure accurate data collection and provide personnel capable of answering questions and conducting interviews with residents who have low levels of literacy to ensure accurate population statistics. All of the personnel that perform these duties need to undergo the appropriate level and type of language tests to ensure that their jobs are being done accurately, making language testing increasingly important to many government agencies.

## Work in Progress

Research into language testing within the US government is largely focused on improving assessment to respond to changing needs in the agency. Language testers in the government produce, administer, and score tests to ensure continued quality results. A typical focus of research in the US government is quality assurance, validity, and efficiency, meaning how to produce results faster or using fewer resources.

In the mid- to late twentieth century, research paid attention to the impact of factors affecting the way the ILRs functioned. Higgs and Clifford (1982) investigated the proportions of rating factors (such as structures and vocabulary) contributing to ILR ratings. Child (1987) outlined the requirements for his ILR-based reading text typology. Lowe (2001) examined the wordings of the ILR Skill Level Descriptions at each level, examining best case, average case, and worst case statements and how these worked for rating in the four proficiency skills. These seminal works were accompanied by others that investigated the nature of the ILR scale and proficiency testing.

The US government's early use of only proficiency exams was based on the fact that most early examinees were native speakers of English and that native speakers of English only need to be tested in receptive skills in the foreign language. Research by Lunde and Brau (2005, 2006) investigated the correlation initially between reading and translation abilities and later between writing and translation abilities. The research found no significant correlation between strong translation ability and strong ability in either reading or writing, leading to the conclusion that a separate skill, the ability to transfer language from one language to another, was needed beyond knowledge of the two languages to successfully translate. In 2015, this research was updated with a larger data set including more languages and the same conclusions were drawn (Brooks and Brau 2015). Consequentially, it is not

advisable to use reading and writing proficiency tests to predict translation ability; translation tests should be administered.

Government language testers utilize hundreds of human raters evaluating a large number of exams, so there is a logical interest in rater reliability and the effects of various rater characteristics, such as native speaker status, rater language proficiency, and rater first language. Rater characteristic research has benefited from studies done within the government context, as it often deals with language proficiencies higher than those typically achieved through academic contexts and with more formalized, large-scale assessment. For example, Brooks (2013) showed how native speaker status has no significant impact on speaking test ratings but rater proficiency level does. The research supported the movement to remove references to the native speaker as a standard for assessment from testing documents and as a requirement for raters.

The importance of standard setting is recognized, and has been most widely used by DLIFLC for the DLPT. Beginning in 2009, the Department of Defense began standard-setting studies to set cut scores according to the ILR Skill Level Descriptions for the DLPT. A standard-setting study engages a panel of language experts who evaluate the item difficulty according to the ILR-SLDs and judge the likelihood of an examinee at a particular level of proficiency to succeed at each item (Impara and Plake 1997). The information provided by the judges, who also have access to pilot test data, is used in the calculation of cut scores for each ILR level. In addition, a larger-scale research effort is underway at the Department of Defense to isolate factors that affect difficulty of understanding audio material, beyond the factors referenced in the ILR Skill Level Descriptions. An initial study on the effect of the density of spoken texts on comprehension is in the planning stages.

The Testing and Assessment Expert Group (TAEG) is a focus group that operates under the Foreign Language Executive Committee (FLEXCOM) of the US Office of the Director of National Intelligence. It is made up of language-testing experts and representatives from various government agencies. TAEG conducted an unpublished interagency comparability study of speaking tests including three agencies and over 150 examinees conducted from 2009 to 2012. As a result of this study, there has been support for annual interagency comparability workshops where the four agencies with speaking test programs (CIA, DLI, FBI, and FSI) meet to review speaking tests and discuss protocol in an effort to better understand each other and norm to the ILR Skill Level Descriptions (Office for the Director of National Intelligence 2016).

## Problems and Difficulties

US government language testers face a constant challenge. On the one hand, they are expected to provide assessments that meet operational demands in critical situations that may arise without warning, and at the same time, they maintain high standards of test validity and score reliability. This combined with the demand to administer thousands of tests annually in an increasing number of languages taxes government resources.

Fluctuating operational needs such as changes in language-related positions, responsibilities, and personnel often call for realignment of test batteries and passing scores or, in many cases, the development of an entirely new test. Often, there is not a large enough population of speakers of the tested language in order to trial the test thoroughly. Test developers must rely on modifying existing test instruments from within their agency or borrowing them from partner agencies. Production time frames by far less than needed for development and validation. Often test development deadlines must be met without additional funds or personnel. Developers rely on in-house technical personnel paired with translators from the field to produce the needed instrument.

The broad range of languages needed and classification of those languages and dialects pose challenges. The US government regularly has a need to communicate or process work in hundreds of languages, representing most language families. Acquiring, training, and evaluating personnel for so many languages pose challenges. Further, many languages have multiple variants or dialects and decisions need to be made as to whether or not it is appropriate to test them separately. Such decisions are often guided by considerations of mutual intelligibility and established recognition of the languages as separate and operational needs; all of these considerations may change with time. For example, Serbo-Croatian was once tested as a single language, but Serbian, Croatian, and Bosnian are now considered independent languages. These decisions are necessary but also costly.

Since the ILR Skill Level Descriptions are used across multiple languages, there are challenges in how to interpret language proficiency equivalently when languages function differently. Issues of diglossia and the acceptability of other "foreign" language features are of issue in language evaluation. Indian subcontinent languages such as Hindi, Punjabi, and Gujarati incorporate a lot of English, and it would at times be incorrect or inappropriate to use the Hindi/Punjabi/Gujarati word in certain contexts even when one exists. Moreover, creoles and patois often convert to other languages when certain proficiency levels are reached. For example, Haitian Creole becomes French for certain functions and contexts. When high-level language functions require shifting to another language, government agencies are challenged to decide whether the upper level functions can be supported by the test language and, therefore, whether or not an examinee can reach the highest level of the scale in that language (Brooks and Mackey 2009).

In Arabic dialects, for example, professional, sophisticated, or contextualized language tasks would never be conducted in the dialect, but rather in Modern Standard Arabic (MSA). It is for this reason that many US government agencies are shifting from testing Arabic dialects in isolation to testing the dialect combined with MSA, particularly in speaking exams. In 2010, the FBI began combining the tests, followed by FSI shortly thereafter. Combined Arabic testing is now being adopted by other agencies. MSA-only tests still exist to evaluate the language of personnel who have taken MSA training courses.

Government language evaluators are challenged to educate the test score users within the organization: the managers, the operational staff that need linguists, and the examinees themselves. Typically, test score users are not accustomed to the

nature of language or are not familiar with the ILR Skill Level Descriptions, leading to confusion, misunderstanding, and inappropriate score use. The indeterminate nature of language, with endless room for interpretation, can lead users to the conclusion that the language test scores are grossly subjective and therefore not accurate. Examinees often misinterpret their ratings' corresponding descriptions to mean the entirety of what a person can do, not the minimum threshold of that level. Likewise, untrained users can misinterpret what a score represents and assign an inappropriate operational task such as giving a translation task to an individual with a high speaking score. To combat this misuse of scores, many US government agencies now provide assessment literacy trainings to examinees and other stakeholders. The trainings are tailored to particular stakeholder audiences to help understand the nature of the ILR scales, how ratings are assigned and how they can be interpreted.

## Future Directions

The focus for government language testing has historically been on producing a useful product that meets the immediate need. Although there have been guidelines for individual tests developed, there have not been set US government standards for quality of language tests or requirements for language-testing procedures; these standards have been left to the individual agencies. With the initiation of the newest generation of DLPTs in 2000, language-testing professionals were being hired by the DLIFLC to support the initiative. The professionalization effort advanced in 2009, when government language testers formed a subcommittee under the American Society for Tests and Materials (ASTM) to write a standard practice for ILR-based language proficiency testing. This standard practice was produced through collaboration between government personnel from many different agencies and private sector language-testing professionals (ASTM 2011).

There are two US government-based organizations that allow for collaboration among agencies with testing programs and needs: the Testing and Assessment Expert Group (TAEG) and the ILR Testing Committee. TAEG is a group formed under the Foreign Language Expert Group of the Office of the Director of National Intelligence. Its membership is composed entirely of government employees who are either language-testing experts or significant language-testing stakeholders. The committee meets monthly to share information and produce official recommendations and cross-agency initiatives. They catalog all the language-testing capacities of the agencies as well as the standards used for test development and quality assurance. Additionally, they have produced recommendations on quality translation assessment and research the comparability of test scores among agencies. Organizations like TAEG are essential to meeting operational needs, as many of the languages that suddenly become critical for an agency's mission are rarely used or assessed in the USA.

The ILR Testing Committee has long been a venue for collaboration and information sharing among government agencies. Its membership is composed not only

of government employees but also of members of academia and industry. The committee has taken on several projects to promote assessment literacy, including understanding the ILR and the development of self-assessment checklists to accompany the ILR Skill Level Descriptions (Interagency Language Roundtable 2016). The ILR Testing Committee has been involved in efforts to clarify and annotate the ILR Skill Level Descriptions for speaking, reading, and listening, to which end there have been several summits involving government and private sector language-testing professionals coming together to discuss the ILR Skill Level Descriptions and articulate a common interpretation of them.

Recent discussions within the TAEG and the ILR Testing Committee have led to a new initiative to revise the four original proficiency skill level descriptions for listening, reading, speaking, and writing. A subcommittee under the ILR Testing Committee has taken on the task of revising the listening descriptions first (Interagency Language Roundtable 2016). The goal of the revisions is not to change the core meaning of each level, which has been in use for over 30 years, but rather to update them, to remove references to antiquated technologies, integrate new modes of communication that have been introduced, clarify and expand upon some of the supporting statement, and remove controversial and difficult to identify concepts, such as the "native speaker" (Brooks 2013).

The top priority of US government assessment is ensuring that government language personnel are qualified to perform the mission of their agencies. US government agencies have a large number of challenges to overcome: developing appropriate language evaluations for an ever-increasing range of languages with minimal resources under strict time constraints for multiple skills, levels, and purposes, all while maintaining a high level of quality. The US government has been a leader in government language testing and has collaborated with government agencies of other countries on language-testing projects. Today, they are still at the forefront of some aspects of testing, working with rarely assessed languages for practical purposes and finding innovative ways to meet operational government needs.

## Cross-References


- ▶ Criteria for Evaluating Language Quality
- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ High-Stakes Tests as De Facto Language Education Policies
- ▶ History of Language Testing
- ▶ Testing Aptitude for Second Language Learning
- ▶ The Common European Framework of Reference (CEFR)


## Related Articles in the Encyclopedia of Language and Education

Sally Magnan: The Role of the National Standards in Second/Foreign Language Education. In Volume: Second and Foreign Language Education

Chantelle Warner: Foreign Language Education in the Context of Institutional Globalization. In Volume: Second and Foreign Language Education

Wayne Wright, Thomas Ricento: Language Policy and Education in the USA. In Volume: Language Policy and Political Issues in Education

## References


ASTM. (2011). F2889-11, Standard Practice for Assessing Language Proficiency, West Conshohocken: ASTM International. www.astm.org

Brau, M. (2013). ILR-based verbatim translation exams. In E. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceeding s of the ALTE Kraków Conference, July 2011* (pp. 333–343). Cambridge: Cambridge University Press.

Brooks, R. L. (2013). Comparing native and non-native raters of US Federal Government speaking tests. Doctoral dissertation. Retrieved from WorldCat Dissertations and Theses. (Accession Order No. 867157336).

Brooks, R. L., & Mackey, B. (2009). When is a bad test better than no test at all? In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 11–23). Cambridge: Cambridge University Press.

Brooks, R. L., & Brau, M. M. (2015, March). Testing the right skill: The misapplication of reading scores as a predictor of translation ability. Paper presented at the Language Testing Research Colloquium, Toronto.

Bureau for International Language Co-ordination (BILC). (2016). *Bureau for International Language Co-ordination*. Resource document. http://natobilc.org. Accessed 18 Aug 2016.

Child, J. R. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations and concepts* (pp. 97–106). Lincolnwood, IL: National Textbook Company.

Child, J. R., Clifford, R., & Lowe, P., Jr. (1991). *Proficiency and performance testing*. Unpublished paper.

Herzog, M. (2003). An overview of the history of the ILR language proficiency skill level descriptions and scale. Resource document. http://govtilr.org/Skills/IRL%20Scale%20History.htm. Accessed 15 Apr 2015.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. Resource document. http://eric.ed.gov/?id=ED210912. Accessed 2 July 2016.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366. doi:10.1111/j.1745-3984.1997.tb00523.x.

Interagency Language Roundtable (ILR). (1985). *Interagency Language Roundtable Skill Level Descriptions: Speaking*. Resource document. http://govtilr.org/Skills/ILRscale2.htm. Accessed 15 Apr 2015.

Interagency Language Roundtable (ILR). (2016). *Interagency Language Roundtable*. Resource document. http://govtilr.org. Accessed 18 Aug 2016.

Jones, R. L., & Spolsky, B. (Eds.). (1975). *Testing language proficiency.* Washington, DC: Center for Applied Linguistics.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal, 28*(2), 136–150.

Lowe, P., Jr. (1985). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. *Foreign Language Proficiency in the Classroom and Beyond, 17*, 9–53.

Lowe, P., Jr. (1988). The unassimilated history. In *Second language proficiency assessment: Current issues* (pp. 11–51).

Lowe, P., Jr. (1998). Keeping the optic constant: A framework of principles for writing and specifying the AEI definitions of language abilities. *Foreign Language Annals, 31*(3), 358–380.

Lowe, P., Jr. (2001). Evidence for the greater ease of use of the ILR language skill level descriptions for speaking. In J. A. Alatis & A.-H. Tan (Eds.), *Georgetown University Roundtable on Languages and Linguistics, 1999* (pp. 24–40). Washington, DC: Georgetown University Press.

Lunde, R. M., & Brau, M. M. (2005, July). *Correlation between reading and translation ability.* Paper presented at the World Congress of Applied Linguistics, Madison.

Lunde, R. M., & Brau, M. M. (2006, June). *Correlation between writing and translation ability.* Paper presented at the American Association of Applied Linguistics, Montreal.

Myron, H. B. (1944). Teaching French to the army. *The French Review, 17*(6), 345–352.

Office for the Director of National Intelligence (ODNI). (2016). *Foreign language*. Resource document. https://www.dni.gov/index.php/about/organization/foreign-language. Accessed 18 Aug 2016.

Petersen, C. R., & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement, 36*, 369–380.

Pulliam, R., & Ich, V. T. (1968). The defense language proficiency tests: Background, present programs, and future plans. *Proceedings of the 10th Annual Conference of the Military Testing Association*, 69–82. Resource document. http://www.internationalmta.org/Documents/1968/Proceedings1968.pdf. Accessed 15 Apr 2015.

# Testing Aptitude for Second Language Learning

Megan Smith and Charles W. Stansfield

**Abstract**

The construct of aptitude for language learning began with the work of John Carroll, who conceived of aptitude as a relatively fixed set of attributes that made some people better able to learn a second language than others. Carroll's work culminated in the Modern Language Aptitude Test, and Carroll's work on aptitude in general and on the MLAT in particular continues to provide the foundation for the development of new aptitude tests and for research and theory related to the role of aptitude in second language acquisition (SLA) research. Since the development of the MLAT, several other aptitude tests have been developed. The US Department of Defense has played a particularly large role in developing aptitude tests, including the Defense Language Aptitude Battery and the High-level Language Aptitude Battery (Hi-LAB). Within SLA research, researchers have refined aptitude as a construct and investigated whether aptitude plays a role in various aspects of second language (L2) learning. One of the ways aptitude has been investigated in SLA research is as an individual factor that might predict ultimate attainment. Some studies find a significant role for aptitude in predicting ultimate attainment – aptitude scores correlate in these studies with end-state L2 knowledge and performance. The present chapter discusses the construct of aptitude, how aptitude is measured, and how it has been used in SLA research.

M. Smith (✉)
Department of English, Mississippi State University, East Lansing, MS, USA
e-mail: megan.smith@msstate.edu

C.W. Stansfield
Language Learning and Testing Foundation, Rockville, MD, USA
e-mail: cstansfield@LLTF.net

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_5

**Keywords**
Aptitude • Modern Language Aptitude Test (MLAT) • High-level Language Aptitude Battery (Hi-LAB) • Second language acquisition • Critical period hypothesis

## Contents



## Introduction

Some people seem to be better able to acquire a second language than other people. The traits that underlie this perceived difference in ability are collectively called aptitude, and the need to identify these individuals has led to the development of several measures of aptitude. Aptitude measures largely developed separately from second language acquisition research, partly because the first aptitude test was developed before second language acquisition emerged as a separate discipline. Thus, although there are obvious overlaps between linguistic aptitude and second language aptitude, it is only relatively recently that second language acquisition researchers have begun to take aptitude seriously as a construct with explanatory power in the study of second language acquisition (SLA). The present chapter discusses several different aptitude tests and also presents an overview of how aptitude has been discussed in second language acquisition research.

## Overview of Aptitude Tests

Linguistic aptitude is thought to be a relatively stable set of attributes that predicts how well an individual can learn a second language (L2). The first aptitude measure, the Modern Language Aptitude Test (MLAT, Carroll and Sapon 1959), was developed in the 1950s with a grant from the Carnegie foundation. Carroll (1971, 1990) identified four relatively independent subcomponents that underlie aptitude: phonetic coding, rote memory, grammatical sensitivity, and inductive reasoning. The MLAT consists of five sections that test phonemic coding ability, rote memory, and grammatical sensitivity. Inductive reasoning ability is not measured. As originally conceived, the MLAT predicts the rate at which people will learn a second language in an instructed setting. Thus, it is not necessarily designed to predict ultimate

attainment, nor is it necessarily designed to identify successful learners in naturalistic language learning contexts. MLAT scores do consistently have moderate to strong correlations with classroom outcomes. Carroll was instrumental in developing the construct of aptitude, and to this day, the MLAT remains the standard against which other aptitude tests are measured (Grigorenko et al. 2000).

In the decades since the development of the MLAT, other aptitude tests have been developed. Within the USA, the Department of Defense (DoD) has had a consistent interest in identifying members of the armed services with high linguistic aptitude in order to select service members for language learning. Consequently, the DoD has encouraged the development of a number of aptitude tests, which include the Defense Language Aptitude Battery (DLAB), and the VORD, which, although it was designed to identify native English speakers with an aptitude for learning non-European languages, has less predictive validity than the MLAT (Parry and Child 1990). Most recently, researchers at the Center for the Advanced Study of Language (CASL) have developed the High-level Language Aptitude Battery (Hi-LAB; Linck et al. 2013). Unlike the MLAT, which predicts language learning at early stages, the Hi-LAB is designed specifically to identify those people who can achieve near-native proficiency in the second language. The DLAB and the Hi-LAB are described in more detail in the next section.

## Aptitude Tests Developed by the Department of Defense

The first test developed by the DoD is the Defense Language Aptitude Battery (DLAB), which was developed by staff at the Defense Language Institute during 1969–1972 (Peterson and Al-Haik 1976). The test requires over 1 h to complete. The DoD has relied on the DLAB to determine aptitude for language learning since it was developed. Military recruits and staff requesting to be assigned to language school must take the DLAB. The DoD has conducted a number of unpublished studies of the construct and predictive validity of the DLAB over the years. The studies have confirmed the predictive validity of the DLAB and DoD staff are satisfied with the test.

Security on the DLAB is high, and there is no official practice test. Still, several commercial study guides have been published, and they can be easily identified through a Google search. Like many such unofficial guides, they make many contradictory statements, even disagreeing on the number of items on the test. Still, they give researchers and potential examinees some idea of the nature of the test. A significant part of the test deals with the learning and application of grammatical concepts in an artificial language. Phonological sensitivity is also tested.

The same DLAB form was administered to recruits showing an interest in being assigned to language school from 1972 to 2008. In 2007, the DoD, through contractors, began developing additional forms and soon afterward began administering them by computer. Specifications were also developed through reverse engineering, since specs for the original form did not exist. The new forms were field tested and equated and are now operational.

In addition to the DLAB, in 2009, DoD contractors developed the Pre-DLAB in order to determine whether prospective language students could pass the DLAB. So while the DLAB is designed to predict success in language training, the short Pre-DLAB is designed to predict success on the longer DLAB. It was found that the Pre-DLAB was able to predict whether an applicant would pass or not pass the DLAB 78% of the time. Multiple forms of the Pre-DLAB were constructed and equated. It should be noted that the Pre-DLAB does not use any item types employed by the DLAB, in order to avoid any potential practice effects between the two tests. In that sense, both tests tap into the language aptitude construct.

The most recent aptitude test developed under the auspices of the DoD is the Hi-LAB (Linck et al. 2013). In the development of the Hi-LAB, the authors developed and piloted a number of cognitive and perceptual measures, administering them to a selection of government employees with language skills at the levels of fluent and advanced (near-native). Prior to developing the tests, the authors surveyed the SLA and cognitive psychology literature to determine what cognitive and perceptual skills are associated with high attainment in a second language. Then they used those findings and personal judgments to develop measures that might be associated with high-level proficiency. The Hi-LAB consists of 11 measures testing executive functioning, memory, and phonemic awareness. Each of these three skills is associated with more than one test, and all tests were delivered on a computer. These measures are described below.

Four tests measure some aspect of executive functioning, which allows people to plan, organize, and complete mental tasks. The *running memory span test* is a comprehensive measure of executive functioning. The RMS focuses on the updating component of executive functioning. In this test, the subject hears 12–20 consonants and must demonstrate recall of the last six in order. Twenty such lists of consonants are presented. The *antisaccade test* is also a measure of executive functioning, but it specifically measures the inhibitory control component. A letter, either B, P, or R, is displayed on different sides of a computer screen for 1/10th of a second along with a visual cue, such as a dot. In order to identify the letter, the subject must not look at the dot, which appears 50 ms prior to the letter. The *Stroop test* also measures the inhibitory control component of executive functioning. The words red, green, or blue appear in a rectangle on a screen. The letters and the rectangle also appear in one of these colors, but the color may not match the meaning of the word. The score represents the reaction time required to identify the color indicated by the word when the letters and rectangle are of a different color. Lastly, the *task-switching numbers test* measures the task-switching component of executive functioning. The test contains two kinds of number identification tasks, which may be presented in sequence or in alternation. Participants receive two response time scores: one for time when the tasks are of the same type and another score for when the tasks are mixed.

The Hi-LAB also includes four subtests that measure some aspect of memory. Two of these subtests, the *letter span test* and the *nonword span test*, both measure working memory and specifically the phonological loop component of working memory. A total of 21 lists of three to nine letters (all consonants) are presented.

Each letter in each list is presented for 0.9 s. The score is the total number of letters the subject recalls in their correct position. In the nonword span test, 15 lists of seven phonotactically plausible one or two syllable nonsense words are presented, each for 2 s. At the end of each list, the subject is shown 14 nonsense words, half of which were on the list. The subject must indicate after each word whether or not it was on the list of words previously shown. Higher scores indicate greater phonological short-term memory capacity. The *paired associates test* is a measure of associative memory adapted from the MLAT. The subject learns 20 pairs of words, an English noun paired with a nonsense word, in a fictitious language. Each pair of words is presented five times for 5 s. During the recall phase, the examinee is presented the nonsense word and must type the English word. The score is the number of correctly recalled English words. The *long-term memory synonym test* consists of a priming task and a comparison task. Participants listen to five words in a list and are then shown two more words, one of which is a synonym for two in the list and the other is a synonym for three words in the list. Participants indicate which word has more synonyms in the list. This exercise alternates with a comparison task in which the subject is presented with pairs of words and must indicate if the words in the pair have similar or different meanings. Some of the pairs contain words that are synonyms of words presented earlier. If the subject recognizes a larger number of words with previously introduced synonyms, they get a positive score. If they are below average in such recognition, they get a negative score.

The serial reaction time test is a measure of sequence learning and implicit learning. An asterisk appears in one of four boxes, and the subject must press the button indicating the box. The subject works through six blocks of 96 responses. Some blocks involve random-like appearances of the asterisk, while other blocks involve a repeating pattern in the location of the asterisk. Response times are totaled under both conditions. Lower scores show faster processing.

Two subtests measure components of test candidates' ability to perceive and discriminate between new sounds. In the phonemic discrimination test, the subject is presented with minimal pairs in Hindi. In each case, the minimal pair involves discriminating between /ǰ/ and /č/, as in the English words /just/ and /church/. Because these phonemes are so similar in Hindi, distinguishing between them often poses a challenge for English speakers. A high score is posited to reflect higher perceptual acuity. The phonemic categorization test is also posited to identify those individuals with higher perceptual acuity. The subject is presented with 90 sounds containing two Russian phonemes that differ only in voicing, such as /d/ and /t/ as in the syllables /da/ and /ta/. Although English also has a voicing contrast between these two phonemes, they have shorter voice onset times in Russian than they do in English, posing a problem for the English speaker.

Researchers at CASL tested government employees and service members whose jobs required them to have high-level language proficiency on the Hi-LAB. In all, 476 subjects participated in the study. All were native speakers of English who had not had significant exposure to a foreign language (such as frequent parental input or immersion abroad) prior to age 10. High-level proficiency was defined as level 4 on the US Government's ILR scale in one language or level 3 or above in three

languages, as demonstrated on the Defense Language Proficiency Test (DLPT). Some subjects were not tested but indicated that they had previously been assigned by a supervisor to work on tasks that required level 3 or 4 skills.

The researchers found that on certain cognitive variables (those involving phonological short-term memory, associative learning, and implicit learning), the subjects who had attained high-level language proficiency scored higher on the Hi-LAB than subjects who had not attained high-level proficiency. Neither the MLAT nor the DLAB were used as comparison aptitude measures in this study, so we do not know if there was any gain in predictive validity for the Hi-LAB over existing tests for the individuals tested. Nonetheless, both the approach and the findings are interesting.

## Other Aptitude Tests

In addition to the MLAT and the tests developed by the DoD, other aptitude tests have been developed. These include the Pimsleur Language Aptitude Battery (PLAB; Pimsleur 1966), the Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test (CANAL-F; Grigorenko et al. 2000), and the LLAMA aptitude test (Meara 2005). The CANAL-F Test operationalizes foreign language learning differently than the MLAT; test candidates learn an artificial language, Ursulu, and their ability to do so is tested with five subcomponents. These test candidates' ability to learn neologisms, to understand the meaning of passages, to learn paired associates, to make selective inferences, and to learn linguistic rules. The first four sections each have two parts: an immediate recall and a delayed recall section. The last section, which tests candidates' ability to learn linguistic rules, has an immediate recall section only. The LLAMA (Meara 2005), in contrast, is largely based on the MLAT. The LLAMA was originally based on the Swansea language aptitude test (SLAT) and has four subcomponents: a test of vocabulary learning, a test of sound recognition, a test of sound-symbol associations, and a test of grammatical inferencing. The LLAMA is available as a free download (www.lognostics.co.uk/tools/llama/index.htm), which makes it accessible to researchers. However, it has not been standardized, and the test developers state that it should not be used as a substitute for the MLAT in high-stakes testing situations. Perhaps because of its accessibility and ease of use, however, the LLAMA is becoming a common aptitude test in SLA research.

Since the development of the MLAT, government and educational institutions have largely used aptitude tests for screening and placement purposes. The US Military uses aptitude measures to place students into language classes at the Defense Language Institute, and the MLAT is used by the defense department of most English-speaking countries (i.e., Canada, the UK, Australia, Singapore, New Zealand) for the same purpose. Many US universities use aptitude measures to identify students with learning disabilities and to provide exemptions from foreign language requirements (e.g., Sparks and Javorsky 1999; Sparks et al. 2002, 2005). Over a two dozen missionary organizations use the MLAT to determine whether two candidate for a mission should be assigned to learn a second language and how

different or difficult the language could be. Similarly, aptitude measures are sometimes used to identify bilinguals who would make good interpreters (e.g., Russo 2011).

While language aptitude tests are mostly used with adults, especially in government, higher education, and business, they are also used but to a lesser extent in other contexts. The use of the MLAT-Elementary (Carroll and Sapon 1967) with children ages 8–11 is growing in the USA but especially in the UK. This may be due to increasing interest in foreign language learning in those countries, as well as the increase in the number of both public and private schools that emphasize foreign language within their curriculum. Interest in determining early whether the student has a FL learning disability is also a reason. The use of the Pimsleur Language Aptitude Battery is increasing somewhat at the secondary level, possibly for the same reason but mainly to determine if a secondary level student has a foreign language learning disability. Nonetheless, in general language aptitude tests are not a part of general education at the elementary and secondary levels. This is probably due to a long-standing concern that a language aptitude test might be misused to exclude students from studying a foreign language when they are truly interested. There is also interest in language aptitude testing in non-English-speaking countries, and while researchers have developed aptitude tests in other languages based on Carrol's theory and approach, these tests are not commercially available, with the exception of a Spanish version of the MLAT-E (Stansfield et al. 2004) and a French version of the MLAT (Wells et al. 1982). Although only used in Canada, the latter could be used in the French-speaking countries of Europe and Africa.

## SLA Research and Aptitude

Since the MLAT was first developed, there has been a significant increase in empirical research that investigates the nature of nonnative language acquisition. Similarly, the predominate approach to foreign language teaching, at least in the USA, has shifted from grammar-translation, which was still dominant in the 1950s, and audiolingualism, which was dominant in the 1960s, to communicative language teaching (CLT). In the early days of SLA research, aptitude received relatively little attention. This is likely due, at least in part, to its association with earlier methods of foreign language teaching, and to Krashen's (1981) distinction between acquisition and learning. Krashen argued that aptitude might predict second language learning but that it should not affect acquisition, which is an unconscious process dependent primarily on interacting with L2 input. As SLA research continued over the years, however, it became clear that SLA is not a uniform process, and researchers began to pay more attention to learner-internal factors and individual differences. As interest in the role that individual differences play in SLA increased, so too did attention to aptitude. Some SLA researchers have argued that aptitude complexes need to be reconceptualized so that they align better with findings from SLA research and to better fit the current models of language teaching (e.g., Robinson 2002; Skehan 2002). Specifically, Skehan (2002) argues that rather than asking whether SLA has

anything to say about aptitude, the question should be whether aptitude subcomponents can shed light on any of the known stages or processes of SLA. For instance, one of the stages that Skehan identifies involves identifying and extending patterns in the L2 input. He argues that the constructs of grammatical sensitivity and inductive reasoning, from the MLAT, are relevant for this stage. In addition to conceptual arguments about the role of aptitude in SLA research, SLA researchers have also investigated whether aptitude effects play a role in important topics in second language acquisition, such as the critical period hypothesis, learners' ability to use feedback, and learning conditions. Each of these will be discussed in turn.

The critical period hypothesis (CPH, Lenneberg 1967) states that the ability to learn a first language (L1) declines as children grow. Although Lenneberg's formulation of the CPH was limited to child L1 acquisition, a significant body of literature (e.g., Abrahamsson and Hyltenstam 2008; Granena and Long 2013; Johnson and Newport 1989, 1991) has investigated whether the age at which people begin acquiring a second language predicts their ultimate attainment for that particular language. A subset of this literature has investigated whether aptitude predicts ultimate attainment in the L2 for people who began learning their L2 after puberty. DeKeyser (2000) tested L1 Hungarian speakers' knowledge of grammatical and ungrammatical English structures using a grammaticality judgment task (GJT). Participants also completed a Hungarian version of the MLAT. Five out of the six participants in DeKeyser's study who had arrived in the USA as adults and received high scores on the GJT also had high aptitude scores. DeKeyser argues that these results suggest that high-level attainment in second language acquisition depends on individuals' analytic ability.

Two recent studies have investigated the relationship between bilingual attainment, L1 maintenance, and aptitude. Abrahamsson and Hyltenstam (2008) investigated the hypothesis that aptitude would play a significant role in acquiring near-native Swedish. They tested 42 native Spanish speakers who spoke Swedish as an L2. These native Spanish speakers had been classified by a group of native Swedish speakers as L1 Swedish speakers. Because age of onset (AO) of L2 learning is hypothesized to predict ultimate attainment, 17 of the participants in this study were late arrivals, having arrived in Sweden after the age of 12. The remaining participants were early arrivals. Participants completed a version of the SLAT that was adapted for Swedish and two GJTs. The participants in the late arrival group had significantly higher aptitude scores than those in the early arrival group, suggesting that aptitude plays a role in achieving nativelikeness for adult language learners. Similarly, Bylund et al. (2012) investigated the role that aptitude played in whether bilinguals were able to maintain native-like L1 proficiency and develop native-like L2 proficiency. In this study, 42 near-native Swedish speakers who spoke Spanish as an L1 completed the SLAT, aural Spanish and Swedish GJTs, and a cloze test that measured grammatical and semantic inferencing skills. The results of this study suggested that there is a close relationship between L1 and L2 proficiency, so that participants who had maintained native-like proficiency in Spanish were more likely to demonstrate native-like proficiency in Swedish. These researchers also found that aptitude, and not amount of daily L1 use or AO, predicted nativelikeness in

bilinguals. Thus, there is at least some evidence that suggests that aptitude scores predict ultimate attainment for adult language learners.

A perennial question in SLA research is that of the role and efficacy of corrective feedback. Although there is debate about whether corrective feedback does anything to make learners' interlanguage systems more target-like (see e.g., Ferris 1999; Truscott 1996), most researchers assume that corrective feedback is beneficial to learners. Thus, researchers have asked what types of feedback learners seem to be able to make use of. Li (2013) investigated the relationship between two aptitude components, analytic ability and working memory (WM), and two feedback types, implicit and explicit feedback. Participants were L1 English speakers learning Chinese as an L2. All participants completed the words in sentences section of the MLAT as a measure of analytic ability and a listening span test as a measure of WM. Participants were then assigned to one of three groups: a control group, an implicit feedback group, and an explicit feedback group. All groups completed two elicited production tasks designed to elicit classifiers. The implicit feedback group received feedback on errors in the form of recasts, and the explicit feedback group received feedback in the form of recasts and an explanation. Participants then completed two GJTs that contained sentences with grammatical and ungrammatical classifier use. The results suggested that analytic ability predicted learners' ability to make use of implicit feedback and that WM scores predicted learners' ability to make use of explicit feedback. In both cases, these effects were found for delayed posttest scores. Li argued that these results suggest that the different task demands required different types of processing, and thus learners' success in each condition was partly a function of their ability to make use of the information they had available to them.

This study highlights an important departure from Carroll's work in recent SLA theory and research. Carroll conceived of aptitude as a relatively fixed attribute. Learners either had high aptitude or they did not, and that was unlikely to change. In recent years, however, SLA researchers (e.g., Robinson 2007; Skehan 2002, 2012) have argued for the existence of aptitude complexes. Aptitude complexes are a set of overlapping traits that collectively underlie language learning ability. These traits are overlapping in the sense that all of them contribute to language learning ability, but an individual could, for instance, have a relatively high working memory capacity and a relatively low analytic ability, and a different individual could have a relatively low working memory capacity and a relatively high analytic ability. These two individuals might have the same aptitude composite score, but have different strengths when it comes to the task of language acquisition. At the same time, some research suggests that MLAT scores are related to L1 verbal ability (Sparks et al. 2011; Sparks 2012) and that exposure to and experience in learning a second language has a positive influence on aptitude scores, indicating that aptitude may not be a fixed trait (Safar and Kormos 2008; Thompson 2013).

Related to this is the question of the relationship between aptitude and medium of instruction. The MLAT was developed and validated with classroom language learners whose language classrooms were either grammar translation or audiolingual in nature. Some research suggests that the MLAT scores do not predict language

learning in communicative classrooms (e.g., Robinson 2007; Safar and Kormos 2008) or in naturalistic settings (e.g., Linck et al. 2013). Two recent experimental studies also failed to find an effect for aptitude under experimental settings. VanPatten et al. (2013) investigated whether grammatical sensitivity scores and explicit information played a role in learners' ability to modify processing behaviors in response to input and feedback. The researchers used the framework of processing instruction (PI) to identify learners' ability to comprehend sentences that violate the First Noun Principle (FNP) in four different languages: Spanish, French, Russian, and German. The FNP states that learners will interpret the first noun or pronoun as the subject or agent of the sentence. L1 English speakers who were enrolled in third semester Spanish, French, Russian, or German classes completed the words in sentences section of the MLAT and a comprehension task designed to alter their processing behaviors. For the comprehension task, learners completed a picture-matching task in which they listened to sentences that conformed to the FNP and did not conform to the FNP. They chose a picture that matched the sentence they heard. Trials to criterion – the number of items it took for participants to start distinguishing between the two sentence types – was used as a measure of rate of acquisition. The researchers found that aptitude scores did not correlate with either trials to criterion scores or posttest scores. Similarly, VanPatten and Smith (2015) investigated whether aptitude predicted L1 English speakers' ability to extend SOV word order to clauses to which they had not been exposed on the basis of limited input. Participants in this study were native English speakers who had had no prior exposure to Japanese. They completed the words in sentences section of the MLAT and a 30-min input treatment task designed to teach them basic word order and vocabulary. They then completed a posttest in which they were tested on their sensitivity to violations of head-final word order in structures to which they had been exposed and to novel structures. Aptitude scores had no relationship to participants' ability to generalize to novel structures. The authors of both of these studies suggest that the difference between this research and the other research that has found aptitude effects is that acquisition is operationalized differently in these studies. In both cases, learners were exposed to meaning-based input and were not explicitly taught rules. Similarly, the participants in both of these studies were tested on their comprehension of the target language and not on their explicit knowledge. This is different from most of the studies reviewed earlier in this section that have measured L2 knowledge by means of GJTs, which are generally thought to tap explicit knowledge (e.g., Ellis 2005). Thus, it could be the case that the correlations between MLAT scores and classroom performance are an indication that the MLAT predicts learners' performance on rule-based language learning, which usually consists of a high proportion of memorization and that the relationship may not be so strong in other instructional contexts.

The renewed interest in aptitude among SLA researchers has generated a wide variety of research. One thing that is clear from this research is that the MLAT has proved remarkably resilient, and it seems to predict outcomes in at least some

domains. At the same time, however, as Winke (2013) pointed out, aptitude as a construct remains relatively undefined. It is also unclear *what* it predicts. To date, the most robust effects for aptitude come from studies that have investigated classroom language learners and have used GJTs or other rule-based tests to assess learners' knowledge of the target language. The jury is out on whether traditional aptitude measures retain their predictive value with respect to outcomes for naturalistic learners or in communicative classrooms.

## Final Remarks

Given the importance of foreign language proficiency to national security interests and the expense of foreign language training, it makes sense that the DoD has a vested interest in aptitude testing. They have invested a significant amount in developing and validating aptitude measures, and these measures do seem to correlate well with classroom-based language learning. This is a finding that holds up, for the most part, in second language acquisition research. A question that remains to be answered, however, is how robust the finding that aptitude scores do not correlate with naturalistic instruction is. This claim comes from several sources: Linck et al. (2013) note that many of the "misses" in their classification model (high attainment learners not classified as high attainment by the model) had learned their languages through "nonstandard" methods, such as missionary works in a country where their language was spoken and had relatively lower education levels. These "nonstandard" language learning experiences are instances of naturalistic language learning, and the primary focus is on communicating with people in the target language. This may mean that these people spend less time with language textbooks learning grammar and vocabulary than the typical language student. This observation is compatible with the results of VanPatten et al. (2013) and VanPatten and Smith (2015), both of which found that aptitude scores did not correlate with participants' performance on an input-based task. This difference between aptitude scores and language learning outcomes in naturalistic versus classroom settings is worth investigating further. In the meantime, aptitude has proved to be robust construct, with predictive validity for both instructed learners and for ultimate attainment. It has also proved useful in a variety of educational and military domains.

## Cross-References


- ▶ Cognitive Aspects of Language Assessment
- ▶ High-Stakes Tests as De Facto Language Education Policies
- ▶ History of Language Testing
- ▶ Language Assessment in the US Government

## Related Articles in the Encyclopedia of Language and Education

Eva Hjörne, Roger Säljö: Categorizing Learners Beyond the Classroom. In Volume: Discourse and Education

Terry Lamb: Knowledge about Language and Learner Autonomy. In Volume: Language Awareness and Multilingualism

Marjolijn H. Verspoor: Cognitive Linguistics and Its Applications to Second Lan guage Teaching. In Volume: Language Awareness and Multilingualism

## References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*, 489–509.

Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2012). Does first language maintenance hamper nativelikeness in a second language? A study of ultimate attainment in early bilinguals. *Studies in Second Language Acquisition, 34*, 215–241.

Carroll, J. B. (1971). *Implications of aptitude test research and psycholinguistic theory for foreign language teaching*. Paper presented at XVIIth International Congress, International Association of Applied Psychology, Liège.

Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. S. Parry & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–29). Englewood Cliffs/ Washington, DC: Prentice Hall/Center for Applied Linguistics.

Carroll, J. B., & Sapon, S. S. (1959). *The modern language aptitude test. Form A*. New York: The Psychological Corporation. Operational test (2002 ed.), Rockville, MD: Language Learning and Testing Foundation.

Carroll, J. B., & Sapon, S. S. (1967). *The modern language aptitude test – Elementary*. New York: The Psychological Corporation. Revised versions (2002, 2010 ed.), Rockville, MD: Language Learning and Testing Foundation.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–533.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141–172.

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1–11.

Granena, G., & Long, M. (2013). Age of onset, length of residence, aptitude and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311–343.

Grigorenko, E., Sternberg, R., & Ehrman, M. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *Modern Language Journal, 84*(3), 390–405.

Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60–99.

Johnson, J. S., & Newport, E. L. (1991). Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language. *Cognition, 39*, 215–258.

Krashen, S. (1981). *Second language acquisition and learning*. Oxford: Pergamon Press.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal, 97*(3), 634–654.

Linck, J., Hughes, M., Campbell, S., Silbert, N., Tare, M., Jackson, S., Smith, B., Bunting, M., & Doughty, C. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63*(3), 530–566.

Meara, P. (2005). LLAMA Language aptitude tests: The manual. http://www.lognostics.co.uk/tools/llama/llama_manual.pdf

Parry, T. S., & Child, J. (1990). Preliminary investigation of the relationship between the VORD, MLAT, and language proficiency. In T. S. Parry & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 30–68). Englewood Cliffs/Washington, DC: Prentice Hall/The Center for Applied Linguistics.

Peterson, C. R., & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery. *Educational and Psychological Measurement, 36*, 369–380.

Pimsleur, P. (1966). *The Pimsleur Language Aptitude Battery.* New York: Harcourt Brace Jovanovich. Operational test, Rockville, MD: Language Learning and Testing Foundation.

Robinson, P. (2002). Learning conditions, aptitude complexes, and SLA: A framework for research and pedagogy. In P. Robinson (Ed.), *Individual differences in instructed language learning* (pp. 113–133). Amsterdam/Philadelphia: John Benjamins.

Robinson, P. (2007). Aptitudes, abilities, contexts, and practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 256–286). New York: Cambridge University Press.

Russo, M. (2011). Aptitude testing over the years. *Interpreting, 13*(1), 5–30.

Safar, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics, 46*, 113–136.

Skehan, P. (2002). Theorizing and updating aptitude. In P. Robinson (Ed.), *Individual differences in instructed language learning* (pp. 69–93). Amsterdam: John Benjamins.

Skehan, P. (2012). Language aptitude. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 381–395). New York: Routledge.

Sparks, R. (2012). Individual differences in L2 learning and long-term L1-L2 relationships. *Language Learning, 62*(Suppl 2), 5–27.

Sparks, R., & Javorsky, J. (1999). Students classified as LD and the college foreign language requirement: Replication and comparison studies. *Journal of Learning Disabilities, 32*(4), 329–349.

Sparks, R., Phillips, L., & Javorsky, J. (2002). Students classified as LD who received course substitutions for the college foreign language requirement: A replication study. *Journal of Learning Disabilities, 35*(6), 428–499.

Sparks, R., Javorsky, J., & Ganschow, L. (2005). Should the Modern Language Aptitude Test be used to determine course substitutions for and waivers of the foreign language requirement? *Foreign Language Annals, 38*(2), 201–210.

Sparks, R., Patton, J., Ganschow, L., & Humbach, N. (2011). Subcomponents of second language aptitude and second language proficiency. *The Modern Language Journal, 95*(2), 253–273.

Stansfield, C. W., Reed, D. J., & Velasco, A. M. (2004). *Prueba de aptitud para lenguas extranjeras-Versión de primaria*. Rockville, MD: Language Learning and Testing Foundation.

Thompson, A. (2013). The interface of language aptitude and multilingualism: Reconsidering the bilingual/multilingual dichotomy. *The Modern Language Journal, 97*(3), 685–701.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*, 327–369.

VanPatten, B., & Smith, M. (2015). Aptitude as grammatical sensitivity and the early stages of learning Japanese as an L2: Parametric variation and case-marking. *Studies in Second Language Acquisition, 37*(1), 135–165.

VanPatten, B., Borst, S., Collopy, E., Qualin, A., & Price, J. (2013). Explicit information, grammatical sensitivity, and the first-noun principle: A cross-linguistic study in processing instruction. *Modern Language Journal, 97*, 506–527.

Wells, W., Wesche, M., & Sarrazin, G. (1982). *Test d'aptitude aux langues vivantes*. Montreal: Institute for Psychological Research.
Winke, P. (2013). An investigation into language aptitude for advanced Chinese language learning. *The Modern Language Journal, 97*(1), 109–130.

# Assessing Multilingual Competence

Alexis A. Lopez, Sultan Turkan, and Danielle Guzman-Orth

**Abstract**

We live in a globalized world in which immigration, transnational relationships, and technological developments continue to create spaces where speakers of many different languages and cultures interact. These forms of multilingual contact have compelled many areas of research to place multilingualism in the spotlight and investigate the dynamics of multilingual interactions, including how speakers negotiate language differences and how speakers develop proficiencies in multiple languages. Also in the spotlight is the need for valid measures of multilingual competence and multilingual practices. In this chapter, we focus on assessing the linguistic competence of multilingual speakers. We center our discussion primarily around the definition of the constructs in multilingual assessments and on different approaches to measure these constructs. We also highlight how the concept of multilingualism and multilingual competence has shifted; discuss current practices in multilingual assessment and works in progress; identify some challenges in the conceptualization, implementation, and interpretation of multilingual assessments; and call attention to areas in need of further research to develop valid multilingual assessments.

**Keywords**

Multilingualism • Multilingual competence • Multilingual practices • Translanguaging

A.A. Lopez (✉) • S. Turkan • D. Guzman-Orth
Educational Testing Service, Princeton, NJ, USA
e-mail: alopez@ets.org; sturkan@ets.org; dguzman-orth@ets.org

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_6

## Contents



## Introduction

As globalization and immigration continue to shape the contemporary landscape, the ability to communicate in multiple languages is increasingly associated with academic and socioeconomic development. In many places around the world, developing and maintaining multilingualism has become a norm at all education levels. Even where multilingualism is absent from official educational policy, students often arrive at school with a repertoire of multiple languages, and most of them are encouraged to learn additional languages, whether through schooling or through interaction with peers outside of school. With the addition of assessments to immigration and schooling policies, measurement of multilingual – or *heteroglossic* – competence has become increasingly common. Yet, current paradigms for assessing multilingual competence lag behind the most recent views of what it means to learn a second language and what it means to know multiple languages (Canagarajah 2006; Shohamy 2013). Multilingualism is generally assessed based on a *monoglossic* view of languages as separate entities, a view that tends to ignore the complex communicative practices of multilinguals and their simultaneous uses of multiple languages (Shohamy 2013). In this chapter, we focus on assessing the linguistic competence of multilingual speakers. We center our discussion primarily around the definition of the constructs in multilingual assessments and on different approaches to measure these constructs. In doing so, we view multilingualism as the unitary linguistic and sociolinguistic ability of individuals to use more than one language in everyday and academic contexts. Addressing this issue is timely as increasing numbers of students with multilingual backgrounds enroll in schools and often have to learn academic content in an environment designed for monolingual (e.g., English only) and actively participate in increasingly complex societies.

## Early Developments

### Defining Multilingualism

Traditionally, most research in the field of bilingualism/multilingualism and second language acquisition has been done from a monolingual (monoglossic) or fractional view (Grosjean 1985). The underlying assumption in a monoglossic perspective is that there is no difference between the language development of monolinguals and multilinguals. As such, multilinguals have access to multiple detached language systems that develop in a linear fashion (Grosjean 1989). Thus, languages, as they reside in the minds of bilingual or multilingual individuals, are treated as separate entities and not as a unified system.

From this perspective, multilinguals are seen as the sum of two or more monolinguals (Grosjean 1989). Thus, monolingual native competence is the ultimate goal to be achieved by multilinguals (Grosjean 1985). Since only native speakers' norms are considered, a monolingual view of multilingualism implies that multilinguals should have full competence, or native-like control, of two or more languages. However, most bilingual or multilingual speakers rarely achieve native-like competence in all their languages (Grosjean 1989).

### Assessing Multilingualism

Historically, most educational and testing contexts have been dominated by a monolingual or monoglossic paradigm in which multilingualism and the multilingual practices have often been ignored (García and Torres-Guevara 2010). The traditional way to measure bilingualism is to assess the two language systems separately from one another and then to compare the results (Hamers and Blanc 2000). Assessments of multilingual competence that reflect a monoglossic perspective "try to account for ultimate native-like proficiency in all the languages" and "assume that the multilingual is the sum of the native-like monolingual competence in each language" (Stavans and Hoffmann 2015, p. 157). Herdina and Jessner (2002) argue that "as long as bilinguals are measured according to monolingual criteria, they appear to be greatly disadvantaged both in linguistic and cognitive terms" (p. 7).

When multilingual competence is assessed using monolingual constructs, test takers are expected to respond exclusively in the target language, even if they may have multiple languages in their repertoire. Test takers' performances are scored using monolingual scoring rubrics, meaning that if they respond using any other language than the target language (either partially or completely), their responses are usually ignored or penalized. The measures used to assess bilinguals are usually the same used to assess monolinguals (Grosjean 1985). Monolingual assessments tend to ignore the different needs that bilinguals have for the two languages and do not take into account that bilinguals use these languages for different purposes, with different speakers, and in different contexts (Grosjean 1989).

## Major Contributions

In this section, we review how the concept of multilingualism and multilingual competence has shifted and discuss current practices in multilingual assessment.

### Redefining Multilingualism

In recent years, many scholars have called for a more holistic view of language, language acquisition, multilingualism, and multilingual development (Block 2003; Lafford 2007). Our global multilingual reality has caused SLA researchers to begin constructing new theoretical paradigms, and greater focus is being placed on the social context and the language learning environment (Block 2003). Lafford (2007) argues for placing more focus on communicative strategies, which may draw upon resources in multiple languages rather than on mastering a particular language. The interaction between language systems is multifaceted and multidirectional, with each language system influencing and being influenced by other language systems (Herdina and Jessner 2002). Thus, speakers in multilingual communities around the world develop proficiencies in two or more languages and learn how to negotiate the relationships between all these languages.

Various scholars point out that there are differences in the linguistic practices of multilinguals and monolinguals (Cenoz and Genesee 1998; Herdina and Jessner 2002). A bilingual "has a specific linguistic configuration characterized by the constant interaction and coexistence of the two languages involved" (Herdina and Jessner 2002, p. 59). Cenoz and Genesse (1998) explain that although monolinguals and multilinguals share the same range of communicative situations, multilinguals possess "a larger linguistic repertoire than monolinguals" (p. 19). Moreover, multilinguals employ two or more languages in interaction in various domains and communities of practice and draw on all their linguistic resources, using one or more languages in the same discourse, or even in the same utterance. The ability that multilinguals have to shuttle between languages is often referred to as *translanguaging* (García and Wei 2014). Translanguaging is defined as an approach to the use of language that considers the language practices of multilinguals as one unified linguistic repertoire rather than as two autonomous, separate language systems (García and Wei 2014). When speakers translanguage, they are able to strategically use their entire linguistic repertoire according to the context and communicative needs (Otheguy et al. 2015).

Given that multilingual competence is not an absolute or invariable state as the languages of multilinguals are in constant flux (Herdina and Jessner 2002), a multicompetent individual is, therefore, an individual with knowledge of an extended and integrated linguistic repertoire and who is able to use the appropriate linguistic resource(s) for an appropriate occasion (Franceschini 2011). Several scholars suggest that there is a close relationship between multilingualism and multicompetence and that the concept of multicompetence might provide insights into the understanding of multilingual competence and multilingual practices

(Franceschini 2011). Multicompetence takes into account the totality of linguistic knowledge in a multilingual's mind to understand how multilinguals use knowledge of multiple languages and how these languages interact in the mind.

## Multilingual Assessment Continuum

Taking a heteroglossic view that supports the stance that the language repertoire of a multilingual operates as a unified system, Shohamy (2011) places multilingual assessments on a continuum. On one end, each language is viewed as a closed and homogenous construct. Although multiple languages may be used in the same assessment in this approach, only responses in the target language are scored. On the other end of the multilingual assessment continuum, we can view all languages as part of an integrated system in which test takers are allowed to mix languages in a dynamic and fluid way, and responses are scored regardless of the language (s) employed, even if mixing occurs within and across utterances (García and Wei 2014). This heteroglossic perspective on assessment of multilingual competence promotes the use of multilingual practices, including language choice, translanguaging, code-switching, and code-mixing. However, most current multilingual competence assessments can be placed on the first end of the multilingual assessment continuum described above.

An example of a monolingual view can be found in the dual language assessments used in early childhood education programs in the United States (Ackerman and Tazi 2015) and in the evaluation of students identified as English Learners (ELs) for special education qualification (Sanchez et al. 2013). In dual language assessment, ELs are assessed in English and their home language (e.g., Spanish). However, the assessments used for this purpose often treat languages as sets of discrete skills that function independently of one another. In other words, languages are assessed, scored, and interpreted separately. For example, students may be penalized for using their home language on a test of English or using English on a test of the home language, even if the underlying meaning shows skill and understanding.

## Works in Progress

As noted with the continuum example presented earlier, the act of assessing multilingual competence can be implemented in a variety of ways. The following section presents several current applications and methods of assessments.

## Current Applications

In the United States, a research study focuses on designing flexible mathematics bilingual assessments to allow students to translanguage whenever needed (Lopez et al. 2014). This means that test takers are given the opportunity to determine when

and how they would like to use their multiple linguistic and semiotic meaning-making resources rather than relying on task directions to tell them when to do so. In these assessments, students can see or listen to the item in English or Spanish, and they can say or write their responses in any language or mix them if needed. Initial investigations of middle school student performance and perceptions indicate that the scores obtained from mathematics assessments that allow the flexible use of all the students' language resources provide meaningful insight to the students' skills in both language and mathematic content domains (Lopez et al. 2014).

Another effort to understand multilingual competence in the US context is found in Sanchez et al.'s (2013) work measuring bilingual students' Cognitive Academic Language Proficiency (CALP) in English and Spanish. The study's goal of measuring CALP in two languages stemmed from a desire to evaluate the process used to assess ELs for learning-based disabilities. Students with a true disability should score low on CALP measures in both English and Spanish rather than showing differences across them. Researchers administered a battery of bilingual assessments to measure students' abilities. Findings indicated that for 10 students, variations of CALP across English and Spanish did emerge, suggesting the critical importance of bilingual assessment to make appropriate special education referral decisions. Other work in the area of ELs at risk for special education referrals focuses on obtaining a more meaningful understanding of the linguistic variability of students who are ELs. Swanson et al. (2011) conducted several longitudinal administrations of a bilingual battery, resulting in multiple waves of bilingual assessment data. This bilingual data was used to estimate latent models of oral language skill across languages, supporting a holistic interpretation of the students' skills across the Spanish and English assessment data instead of a monolingual interpretation for each language. The modeling yielded four consistent profiles of high performing bilinguals, average performing bilinguals, low performing bilinguals, and bilinguals who were English dominant, suggesting that advanced statistical modeling techniques may help with the meaningful interpretation of the initial underlying nature of natural performance heterogeneity within the EL subgroup (Guzman-Orth and Nylund-Gibson 2013).

## Flexible Multilingual Assessment Methods

Recently, there have been a few efforts to develop flexible multilingual assessment methods that give test takers the freedom to translanguage whenever needed (e.g., Gorter 2014; Lopez et al. 2014). By this, we mean developing multilingual assessments that allow test takers to use whatever features they have in their integrated language system to demonstrate what they know and are able to do with language (García and Wei 2014; Otheguy et al. 2015; Shohamy 2011). Multilingual speakers' linguistic repertoire includes both standard and vernacular varieties (Sayer 2013). If multilingual speakers are not permitted to draw upon their diverse linguistic repertoire, they may be unfairly disadvantaged because the assessment does not allow them to fully display their language skills (García and Wei 2014; Otheguy et al. 2015).

Lopez et al. (2014) proposed a technology-enhanced assessment platform that allows multilingual speakers to use multiple assessment features (e.g., see and listen to item in multiple languages, write or record responses) so they can strategically use whichever languages and language practices they have at their disposal. In this platform, questions are posed in multiple languages from which test takers can choose and they can use all their languages (both standard and vernacular varieties) to answer them. Test takers are also free to mix languages if needed without being penalized. Moreover, test takers can also use different semiotic meaning-making features, enabling them to perform in writing, orally or graphically (Wei 2011). Test takers' responses are scored using conceptual scoring, a scoring method that allows for the scoring without regard to the language or mode in which the response is given (Barrueco et al. 2012). Depending on the language skill being measured, some assessment features can be disabled. For example, if writing is the construct being measured, then test takers will not be able to record their response.

## Problems and Difficulties

The assessment of multilingualism, in its current definition and conception, poses a number of challenges. First, we highlight some problems related to the multiple influences that impact multilingual assessments. Then we describe other challenges from a validity perspective. We will discuss challenges in the conceptualization of the construct, the implementation of multilingual assessments, and scoring and interpretation issues.

### Imposing Language Policies That Neglect Multilingual Diversity

According to Stavans and Hoffmann (2015), "measures of multilingualism are usually driven by educational, political and economic forces rather than socio-psychological ones" (p. 157). In fact, Shohamy (2011) argues that tests serve as tools to impose monoglossic political ideologies to maintain "national and collective identities" (p. 420). This is clearly evident in countries with high numbers of immigrants. For example, language tests are increasingly used in making decisions about immigration and citizenship (McNamara and Shohamy 2008). Many countries now require immigrants to demonstrate proficiency in the dominant (official or national) language to gain residency and citizenship (McNamara et al. 2015). The use of language tests for this purpose typically takes a monolingual or fractional perspective on multilingualism as they tend to not recognize or value the minority languages and language practices of multilinguals (Barni 2015; McNamara and Shohamy 2008). It has been argued that most assessments in the context of immigration and citizenship impose monolingual policies and suppress multilingual diversity by ignoring the overall language competence of immigrants (Barni 2015).

Another example of a monoglossic language policy is the No Child Left Behind Act (NCLB) of 2001 in the United States (2002). NCLB requires all students, including students who have recently immigrated to the United States, to participate in statewide academic assessments for accountability purposes. These academic assessments reflect a monolingual view in the sense that all students, including ELs born in the United States and immigrant students, are required to demonstrate academic proficiency in English. Thus, many immigrant students are deprived of the opportunity to demonstrate knowledge in academic content areas because their English language skills are not fully developed. Current monolingual academic assessments tend to ignore some of the language, knowledge, and experiences that immigrant students bring to school (Lopez et al. 2014; Shohamy 2011). These assessments usually overlook the common practice of immigrant students using their home languages in different academic contexts. Additionally, NCLB holds states accountable for students' progress in English language proficiency attainment. This means that only the students' English language development is valued while proficiency in minority languages and multilingual practices are often ignored from the federal accountability perspective.

## Conceptualizing, Implementing, and Interpreting Multilingual Assessments

One of the biggest challenges in multilingual assessment is conceptualizing the constructs that need to be measured. Multilingual assessments should reflect language practices that are dynamic and fluid and, thus, allow test takers to select language features from their linguistic repertoire in ways that fit their communicative needs (García and Wei 2014; Lopez et al. 2014). To do this, a *paradigm shift* – from a monolingual/monoglossic/fractional view to a multilingual/multiglossic/holistic view – is needed (Shohamy 2013). This requires making changes in assessment policies and practices to promote and value multilingualism. Change is also needed in the implementation and operationalization of the constructs of multilingual assessments. Thus, it is important to develop language standards that are based on a holistic view of multilingual competence, that reflect the complex language practices of speakers in multilingual societies, and that clearly describe linguistic performance in different languages and across languages.

An additional related challenge is finding ways to implement the holistic view of language in multilingual assessments. If the intent is to develop multilingual assessment policies and practices that allow test takers to use their entire linguistic repertoires by accepting and encouraging the mixing of languages, the role of the test administrator becomes crucial. In this type of multilingual assessment, test administrators become mediators in the sense that, along with the test taker, they will work together to negotiate and create meaning. Consequently, one practical constraint arises. Multilingual assessments may require test administrators to possess the same languages and regional dialects as the test takers and be familiar with the

communicative practices and strategies that test takers use to negotiate language differences.

Scoring is also a challenge for multilingual assessments. A holistic view of language defines performance in two or more languages as complementary, in that multilinguals could dynamically use varying language skills, depending on the context and audience. Therefore, appropriate scoring models have to be developed to accommodate this construct definition. Score interpretation is also a challenge in multilingual assessment, given that it is difficult to include every possible target language use situation that multilinguals are expected to engage in within one assessment. Thus, it is critical to examine the specific language skills and functions that may be generalizable from one communicative task to another and from one language to another.

## Future Directions

We live in a globalized world in which immigration, transnational relationships, and technological developments continue to create spaces where speakers of many different languages and cultures interact. These forms of multilingual contact have compelled many areas of research to place multilingualism in the spotlight and investigate the dynamics of multilingual interactions, including how speakers negotiate language differences and how speakers develop proficiencies in multiple languages. The field of language testing will be increasingly compelled to participate in this conversation and to devise valid measures of multilingual competence.

A concept that could help in understanding multilingual competence and multilingual practices is *plurilingual* and *pluricultural* competence as defined by the Common European Framework of Reference (CEFR) because it describes the complex language practices of multilinguals in a more accurate way (Garcia et al. 2007). According to the CEFR, plurilingual and pluricultural competence refers to "the ability to use languages for the purpose of communication and to take part in intercultural interaction, where a person, viewed as a social agent has proficiency, of varying degrees, in several languages and experience of several cultures" (Council of Europe 2001, p. 168).

The Language Policy Division of the Council of Europe developed a guide to suggest how plurilingualism can be promoted in Europe (Beacco and Byram 2007). This guide evaluates language education policies and examines how they can be used to enable individuals to become plurilinguals. Now there are language education programs that aim to develop plurilingual and intercultural competence. For example, the European Center for Modern Languages proposed a framework of reference for pluralistic approaches to languages and cultures (FREPA) (Candelier et al. 2012). A pluralistic approach to teaching languages and cultures is any didactic approach that involves and values more than one language and culture. The goal of the FREPA project is to support the development of plurilingual and intercultural competence of learning at all levels.

Plurilinguals use their entire linguistic repertoires, which is part of a multiple competence, to carry out different tasks; that is, speakers use all their languages

depending on the communicative need. Consequently, assessments of multilingual competence should also assess the learners' abilities to use their entire linguistic repertoires by allowing test takers to use different languages in different situations, for different purposes, and with different people. That is, multilingual assessments should allow test takers to use dynamic and fluid language practices (i.e., translanguaging) (García and Wei 2014; Otheguy et al. 2015).

Although there is a clear need for developing assessments of multilingual competence that reflect a holistic or heteroglossic perspective, more research is needed to develop valid multilingual assessments. For example, the constructs to be measured in multilingual assessments must be clearly defined. Empirical and operational assessment development work should examine the extent to which the multilingualism construct, operationalized through a heteroglossic or holistic view, can be feasibly assessed. Moreover, it is equally important to have a clear understanding of how multilingual communication works. In order to account for multilingualism in assessment, it is also important to examine how multilinguals use their knowledge of multiple languages as well as the communicative practices they use to negotiate meaning in different communicative situations and the strategies speakers use to negotiate their language differences. Thorough analyses of target language use domains in multiple languages and across languages need to be conducted in order to generalize student performance on multilingual assessments to other possible multilingual contexts. Also, scoring models must be designed and validated in accordance with a holistic view of multilingualism. Equally important is understanding the various purposes and uses for multilingual assessments (e.g., for instructional or accountability purposes). Who should be tested? In what contexts? How should the results be interpreted and used, and by whom? Addressing these areas of research will require combined efforts across different but related disciplines, including SLA, multilingualism, psycholinguistics, sociolinguistics, and language testing.

## Cross-References


- ▶ Assessing English as a Lingua Franca
- ▶ Critical Language Testing
- ▶ Utilizing Accommodations in Assessment


## Related Articles in the Encyclopedia of Language and Education


Adrian Blackledge, Angela Creese: Language Education and Multilingualism. In Volume: Language Policy and Political Issues in Education

Jasone Cenoz, Durk Gorter: Translanguaging as a Pedagogical Tool in Multilingual Education. In Volume: Language Awareness and Multilingualism

Ofelia García: Extending Understandings of Bilingual and Multilingual Education. In Volume: Bilingual and Multilingual Education

Ulrike Jessner-Schmid: Multicompetence Approaches to Language Proficiency Development in Multilingual Education. In Volume: Bilingual and Multilingual Education

Olga Kagan, Kathleen Dillon: Issues in Heritage Language Learning in the United States. In Volume: Second and Foreign Language Education

Beatriz Lado; Cristina Sanz: Methods in Multilingualism Research. In Volume: Research Methods in Language and Education

Leslie Moore: Multilingual Socialization and Education in Non-Western Settings. In Volume: Language Socialization

Ingrid de Saint-Georges: Researching Media, Multilingualism, and Education. In Volume: Research Methods in Language and Education

Massimiliano Spotti, Sjaak Kroon: Multilingual Classrooms at Times of Superdiversity. In Volume: Discourse and Education

## References


Ackerman, D. J., & Tazi, Z. (2015). *Enhancing young Hispanic dual language learners' achievement: Exploring strategies and addressing challenges* (Policy Information Report No. RR-15-01). Princeton: Educational Testing Service.

Barni, M. (2015). In the name of the CEFR: Individuals and standards. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 40–51). New York: Routledge.

Barrueco, S., López, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish-English bilingual preschoolers: A guide to best approaches and measures*. Baltimore: Paul H. Brookes.

Beacco, J. C., & Byram, M. (2007). *From linguistic diversity to plurilingualeducation: Guide for the development of language education policies in Europe*. Strasbourg: Council of Europe. Retrieved from http://www.coe.int/t/dg4/linguistic/Guide_niveau3_EN.asp#TopOfPage

Block, D. (2003). *The social turn in second language acquisition*. Washington, DC: Georgetown University Press.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3*(3), 229–242.

Candelier, M. (Coord.), Camilleri Grima, A., Castellotti, V., de Pietro, J. F., Lőrincz, I., Meißner, F. J., Schröder-Sura, A., Noguerol, A., &Molinié, M. (2012). *FREPA – A framework of reference for pluralistic approaches to languages and cultures: Competences and Resources*. Strasbourg: Council of Europe. Retrieved from http://www.ecml.at/tabid/277/PublicationID/82/Default.aspx

Cenoz, J., & Genesee, F. (1998). Psycholinguistic perspectives on multilingualism and multilingual education. In J. Cenoz & F. Genesse (Eds.), *Beyond bilingualism: Multilingualism and multilingual education* (pp. 16–32). Clevedon: Multilingual Matters.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Franceschini, R. (2011). Multilingualism and multicompetence: A conceptual view. *The Modern Language Journal, 95*(3), 344–355.

García, O., & Torres-Guevara, R. (2010). Monoglossic ideologies and language policies in the education of U.S. Latinas/os. In E. Murillo, S. Villenas, R. Trinidad Galván, J. Sánchez Muñoz, C. Martínez, & M. Machado-Casas (Eds.), *Handbook of Latinos and education: Research, theory and practice* (pp. 182–194). Mahwah: Lawrence Erlbaum.

García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. London: Palgrave Macmillan Pivot.

Garcia, O., Bartlett, L., & Kleifgen, J. (2007). From biliteracyto pluriliteracies. In P. Auer & L. Wei (Eds.), *Handbook on multilingualism and multilingual communication* (Vol. 5, pp. 207–228). Berlin: Mouton de Gruyter.

Gorter, D. (2014, March). *Approaches to multilingual assessment in education in the Basque country.* Paper presented at the Invited Colloquium, Negotiating the Complexities of Multilingual Assessment, at the annual American Association for Applied Linguistics (AAAL) Conference. Portland.

Grosjean, F. (1985). The bilingual as a competent but specific speaker-hearer. *Journal of Multilingual and Multicultural Development, 6*, 467–477.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*, 3–15.

Guzman-Orth, D., & Nylund-Gibson, K. (2013, April). *When is an English language learner not an English language learner? Exploring individual differences in developmental language and literacy acquisition for at-risk learners: A latent transition approach.* Paper presented at the annual meeting of American Educational Research Association (AERA), San Francisco.

Hamers, J. F., & Blanc, M. H. A. (2000). *Bilinguality and bilingualism* (2nd ed.). Cambridge: Cambridge University Press.

Herdina, P., & Jessner, U. (2002). *A dynamic model of multilingualism: Perspectives of change in psycholinguistics*. Clevedon: Multilingual Matters.

Lafford, B. (2007). Second language acquisition reconceptualized? The impact of Firth & Wagner. *The Modern Language Journal, 91*, 735–756.

Lopez, A. A., Guzman-Orth, D., &Turkan, S. (2014, March). *A study on the use of translanguaging to assess the content knowledge of emergent bilingual students.* Paper presented at the Invited Colloquium, Negotiating the Complexities of Multilingual Assessment, at the annual American Association for Applied Linguistics (AAAL) Conference. Portland.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89–95.

McNamara, T., Khan, K., & Frost, K. (2015). Language tests for residency and citizenship and the conferring of individuality. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 11–22). New York: Routledge.

No Child Left Behind (NCLB) Act of 2001. (2002). Title IX U.S.C. § 9101. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg107.html#sec9101

Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied LinguisticsReview, 6*(3), 281–307.

Sanchez, S. V., Rodriguez, B. J., Soto-Huerta, M. E., Castro Villareal, F., Guerra, N. S., & Bustos Flores, B. (2013). A case for multidimensional bilingual assessment. *Language Assessment Quarterly, 10*, 160–177.

Sayer, P. (2013). Translanguaging, TexMex, ad bilingual pedagogy: Emergent bilinguals learning through the vernacular. *TESOL Quarterly, 47*(1), 63–88.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal, 95*(3), 418–429.

Shohamy, E. (2013). Expanding the construct of language testing with regards to language varieties and multilingualism. In D. Tsagari, S. Papadima-Sophocleous, & S. loannou-Georgiou (Eds.), *International experience in language testing and assessment: Selected papers in memory of PavlosPavlou* (Language Testing and Evaluation Series, Vol. 28, pp. 17–32). Frankfurt am Main: Peter Lang.

Stavans, A., & Hoffmann, C. (2015). *Multilingualism*. Cambridge: Cambridge University Press.

Swanson, H. L., Orosco, M. J., Lussier, C. M., Gerber, M. M., & Guzman-Orth, D. A. (2011). The influence of working memory and phonological processing on English language learner children's bilingual reading and language acquisition. *Journal of Educational Psychology, 103*(4), 838–856.

Wei, L. (2011). Moment analysis and translanguaging space: Discursive construction of identities by multilingual Chinese youth in Britain. *Journal of Pragmatics, 43*, 1222–1235.

# Assessing English as a Lingua Franca

Jennifer Jenkins and Constant Leung

**Abstract**

English as a Lingua Franca refers to English used as a contact language among speakers of different first languages, whether from choice or through some kind of coercion. English as a Lingua Franca (henceforth ELF) has the largest number of users of English worldwide, of whom the vast majority are nonnative speakers. The resulting diversity and emergent nature of ELF communication mean that it is not amenable to being captured in descriptions of static norms, and therefore that conventional language assessment is ill-equipped to deal with it. This chapter is thus different from the others in the volume to the extent that as tests of ELF do not currently exist, the discussion is primarily conceptual, exploring developments in thinking about assessing ELF rather than contributing to and critiquing specific test types, goals, and descriptors. After introducing the ELF background, we go on to discuss early orientations to assessing nonnative uses of English, in particular, the work done within the related field of World Englishes. We then consider some key conceptual approaches from ELF scholars, both earlier and more recent, along with the influences on these approaches from the field of critical language assessment and research into multilingualism, particularly translanguaging. Next, we discuss the kinds of difficulties that are faced by anyone attempting to pursue a nontraditional ELF-oriented approach to the assessing of English. We end by exploring some possible future directions that ELF-oriented language assessment scholars might take in designing and introducing a radical new way of assessing of English in its ELF guise.

J. Jenkins (✉)
Department of Modern Languages, University of Southampton, Southampton, UK
e-mail: j.jenkins@soton.ac.uk

C. Leung
Centre for Language, Discourse and Communication, School of Education, Communication and Society, King's College London, London, UK
e-mail: constant.leung@kcl.ac.uk

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_7

**Keywords**
ELF • Emergent language • Linguistic diversity • Translanguaging • Language models

## Contents



## Introduction

The term, ELF, refers to the use of English in intercultural communication among English users from any part of the world. In other words, it can involve speakers from any of the postcolonial regions (Kachru's "outer circle"), the Anglophone countries (Kachru's "inner circle," whose varieties he described as "norm-providing"), and the many countries where English is learnt and used by its speakers but has few or no internal functions (Kachru's "expanding circle," which he described as "norm-dependent," i.e., dependent on the norms of Anglophone varieties and later also on outer circle norms; see Kachru 1992, 2005). The crucial point is that we are talking about a use of English that transcends national/first language boundaries, by contrast with the established nativized or developing varieties of English used within any one country of the outer circle. However, while ELF includes the use of English across all these countries, its largest number of users by far comes from the expanding circle, whose English speakers tend to use the language exclusively for international/intercultural (i.e., ELF) communication.

A number of forms have been identified in research as occurring across ELF speakers from a large number of different L1s. These include countable use of nouns that are (currently) uncountable in native English (e.g., "feedbacks," "softwares"), interchangeable use of the relative pronouns "who" and "which," and alternative ways of pronouncing the voiceless and voiced dental fricative "th" (see Seidlhofer 2011 for further examples and discussion). ELF use also involves L1-specific features carried over from speakers' L1s, a phenomenon traditionally described as "L1 transfer." The situation was further complexified by the later recognition that much ELF communication consists of transient interactions co-constructed "online" as speakers from diverse language backgrounds convey and negotiate meaning through accommodation strategies and the like. The English that results is unpredictable, and often characterized by ad hoc, nonce, and hybrid forms. Indeed, the most defining feature of ELF use is now seen as its "variably variable" nature. Despite these potential problems, however, ELF communication is not only

frequent, but also, according to a large body of research, highly successful. This has implications for English language assessment.

## Early Developments

Research into ELF communication, and therefore thinking about assessing ELF, began relatively recently. However, some of the key issues had already begun being tackled within the field of World Englishes (henceforth WE). And as WE research had a major impact on early conceptualizations of ELF, we include the relevant (for ELF) research on testing WE in this section on "Early Developments."

One of the first WE researchers to observe that there was (and remains) a disconnect between the way English is used in context as a social (including professional) practice by nonnative speakers of WE and the Anglophone norms preferred by the major international examination boards was Lowenberg (e.g., 1993). He argued that there was "an implicit, and frequently explicit, assumption . . . that the universal target for proficiency in Standard English around the world is the set of norms which are accepted and used by highly educated *native speakers* of English" (2000, p. 67; his italics). This, as Davidson (2006) points out, called into question the validity of the international English language tests, for although such items might be correct in Anglophone (typically Australian, UK or US) contexts, they were often not appropriate, or even correct, in other local contexts. In addition, as Lowenberg also pointed out, "by not reflecting the sociopolitical reality of non-native varieties, [the tests] may unfairly discriminate against speakers of these varieties" (2000, p. 69).

Lowenberg supported his argument with a range of examples. These demonstrated both non-Anglophone use that was considered correct in the local context, and how, in certain cases, local non-native English use was simply extending a process that was already in place in standard native English. A case in point is the conversion of uncountable nouns to countable. This is a relatively common process in native English but is considered an error in nonnative English unless it has first been "sanctioned" by native English speakers. Examples provided by Lowenberg include countable use of "furnitures," "luggages," and "equipments."

While supporting the use of local norms for the postcolonial Englishes, Lowenberg (2000) had argued in favor of retaining native English norms for the expanding circle Englishes. Two years later, however, he revised his position and extended his argument of the right to local norms to the expanding circle, in other words, to English used primarily for intercultural communication, or ELF (although Lowenberg did not himself use the term "ELF"). His claim was similar to the one he had previously made in respect of WE contexts, if a little more tentative. That is, that "in many of these Expanding Circle settings, the norms for Standard English usage, teaching, and testing may not always be those of the Inner Circle." He also observed that as with outer circle Englishes, "Expanding Circle norms result from productive processes that also occur in the Inner Circle varieties" (2002, p. 431).

Lowenberg now argued, therefore, as he had done previously in respect of the outer circle Englishes, that developments in the expanding circle, too, called into question the validity of international tests predicated on native English. Again, he provided examples to illustrate his claim: morphological and syntactic innovations such as article use (e.g., Korean article use of "a hard work," and "a great patience," p. 432), use of native English uncountable nouns as count nouns, and differences in the formation of phrasal verbs. Lowenberg saw such items as "varietal" features, and thus as "differences" from native English rather than "deficiencies" by comparison with it. This meant, he claimed, that it would be possible to describe expanding circle varieties and distinguish their features from deficiencies caused by faulty second language acquisition. These varietal features, he argued, could then be taken account of, rather than penalized, in international tests.

The following year, Davies et al. (2003) noted, like Lowenberg, that imposing native English norms on nonnative speakers in tests such as TOEIC, TOEFL, and IELTS, and penalizing them for not using these norms, risked being "discriminatory" (p. 571). A little later, Davidson (2006) made a similar point, also repeating a crucial question left open by Davies et al. (2003) as to whose norms should be used in tests of English. Davidson called for more empirical research into the quality and quantity of variation across native and nonnative Englishes to help answer the question, arguing that "large testing companies … will act and act most profoundly when confronted with hard, cold numbers" (p. 714).

It was at this point, i.e., 2006, that ELF research made its first entry into the English language assessment debate. This took the form of a commissioned "Point and counterpoint" between Jenkins and (Taylor 2006) for the 60th anniversary issue of *ELT Journal*. In essence, (Jenkins's 2006) point was that "recent changes in both users and uses of English have become so far-reaching that a major rethink of English language teaching goals is called for" and "that this will first require a substantial overhaul of English language testing" (p. 42). Focusing on lingua franca, i.e., international uses of English, the author's approach was nevertheless that of "early ELF," or what I later called "ELF 1" (see Jenkins 2015). That is, it still looked at ELF from a WE "varieties" perspective. Even so, it already highlighted the in situ co-construction of meaning among nonnative English speakers from different L1s through their use of accommodation strategies and ad hoc creativity. This presaged the forthcoming recognition of variability as the defining feature of ELF communication and a move away from a monolithic framing of language competence in terms of native speaker norms and practices.

Although Taylor, perhaps mindful of her role as a representative of one of the major international language examination boards, scarcely engaged with Jenkins's central argument and seemed to maintain a position that continued to privilege native speaker norms, she nevertheless made an important point in observing that testing is "the art of the possible" (p. 58), and pointing out that recent global changes in English were undoubtedly making the work of the examination boards more difficult. This suggested that Taylor did at least recognize that things could not stay the same and that changes would eventually have to be made. The essence of this debate

and the nascent recognition of the shape of things to come have been borne out by more recent work.

## Major Contributions

As the amount of available empirical evidence from ELF research increased during the first decade of the twenty-first century, ELF's variable, emergent nature became increasingly clear, and the earlier focus on common ELF forms gave way later in the second half of the decade to the understanding, noted above, that contingent variability was ELF's defining feature. This, in turn, meant that ELF could no longer be seen as similar to WE and approached from a varieties perspective. The implications for English language assessment were substantial: if ELF were not a stable variety as such, then there would be no normative references in terms of language forms and/or use for testing, and if this was so, it would be impossible to assess ELF by conventional psychometrically oriented standardized tests of the kind that were currently being administered around the world. In a fundamental way we are drawing on insights from a Hymesian perspective. Blommaert (2015, pp. 21–22) argues that use of language should not be understood "... by reference to 'Language' with a capital L [e.g., English]... but by reference to repertoires. Such repertoires are an organized complex of specific resources such as varieties, modes, genres, registers and styles ... repertoires can only be understood by attending to their functions, i.e., to their actual and contextual deployment, not to any abstract or a priori assessment of what they mean or of what they are worth ...."

However, the major international English language examinations showed, and still show, no inclination to take ELF communication into account in their test design. Instead, they continue to assess candidates' ability with reference to putative native English norms as if they would only be communicating with native English speakers, or nonnative English speakers who only regard standard native varieties as acceptable. The findings of empirical ELF research have thus had no influence to date on the goals of English language assessment and the kinds of English that the boards specify in their descriptors and accept as "correct." In this section, with no tests of ELF available for discussion and examination, we therefore explore the major contributions to the debate around the need for ELF rather than EFL (i.e., native English) to be the main focus of international English language testing. We start by considering two closely linked bodies of research and discussion that have had major influences on thinking about assessing ELF, and in some cases, addressing it directly: critical language testing and multilingualism. We then go on to discuss contributions from inside ELF research.

Work on critical language testing has been critiquing the prevailing monolingual approach of language testing for some time. While the researchers discussed above made their case against the testing of native English from a WE perspective, others have taken a critical view from the perspective of dynamic bilingualism and the notion of the emergent bilingual (Flores and Schissel 2014). Earlier contributions with clear implications for assessing English from an ELF perspective include

Shohamy (2001, 2006), Canagarajah (2006), and Elder and Davies (2006). Three of these date, like Davidson's and the "point and counterpoint" discussed above, from 2006. Shohamy (2006) discusses six common mechanisms that she sees as affecting language policy, one of which is language tests. She points out that in focusing only on native English, the international English language tests are suppressing diversity. In effect, then, by accepting only certain local varieties of native English, i.e., so-called standard American and British English, these tests deny the very "international" character they claim to represent. Meanwhile, Canagarajah (2006) argues against the testing of individual varieties per se, be they native English or WE, and in favor of a more heterogeneous approach with "multiple norms and diverse grammars" and a focus on "performance and pragmatics" (p. 232). Elder and Davies (2006), although not themselves ELF researchers, focus specifically on ELF. They propose two alternative models for assessing ELF. The first entails minor modifications to existing tests to render them "accessible and fair for ELF users without changing the construct" (p. 282). The second involves treating ELF "as a code in its own right," similar to WE (ibid.). Neither of these would be considered acceptable ELF-oriented solution 10 years later, the first because it remains predicated largely on native English, the second because it takes the later discarded (for ELF) "varieties" approach. However, for those coming from an ELF perspective, they are a step in the right direction by at least recognizing that the phenomenon of ELF calls for changes in the way English is assessed internationally.

More recently, another critical language assessment scholar, McNamara has also contributed to the discussion, in three separate articles all of which engage directly with ELF. In the first article, McNamara argues that "we are at a moment of very significant change, the sort of change that only comes along once in a generation or longer – the challenge that is emerging in our developing understanding of what is involved in ELF communication" (2011, p. 507). He goes on to observe that ELF is "a key feature of a globalized world" and as such "presents a powerful challenge to assumptions about the authority of the native [English] speaker, an authority which is enshrined in test constructs" (p. 513). His main focus, however, and again in his 2012 article, is on the Common European Framework of Reference (CEFR), and as the CEFR is discussed in more detail below in relation to ELF, we will return to McNamara (2011) later.

McNamara's 2014 publication is part of a set of articles exploring communicative language testing over the past 30 years, and considering whether it represented "evolution or revolution." McNamara's conclusion is that communicative language testing represents evolution rather than revolution and that it needs fundamental change in two respects, firstly in relation to recent technological advances, and secondly, in terms of the reality of English as it is currently used in lingua franca communication around the world. In the latter respect, he remarks that:

> the growing awareness of the nature of English as a lingua franca communication overturns all the givens of the communicative movement as it has developed over the last 30 or 40 years. The distinction between native and non-native speaker competence, which lies at the heart of the movement, can no longer be sustained; we need a radical reconceptualization

> of the construct of successful communication that does not depend on this distinction (p. 231).

He concludes by arguing that evolution is not sufficient in communicative language testing and calls for revolution or, rather, "revolución!" (ibid.).

We note that McNamara's central argument would apply in other contexts. What counts as native speaker competence is now a moot point even in so-called native English-speaking environments. For instance, Leung and Street (2014) report that in a London school where over 80% of the students were from ethnically and linguistically diverse communities, teacher-student talk in the classroom included not only teaching-learning oriented content-based exchanges but also playful mock *ad hominem* insults that seemed to (re-)affirm their cordial relationship. The intricate weaving of formal pedagogic and informal social talk requires all interlocutors to have a highly tuned sensibility to a local language practice, the maintenance of which requires subtle negotiation of role boundaries and individual tolerances.

Shohamy, too, has engaged directly with issues relating to the assessing of ELF specifically. This took the form of an unpublished plenary paper at the 7th International ELF conference (Athens, September 2014), with the title "Critical language testing and English lingua franca: how can one help the other?" In her plenary, Shohamy argued that ELF, along with the phenomena of translanguaging and bi-multi-languaging, challenges traditional language testing. She went on to observe that "for most people in the world, L2 is viewed as ELF, multilingual, and multimodal" resulting in "new and creative mixes." These mixes, she noted, however, are ignored in English language tests, which continue to impose "monolingual practices" and penalize L1 use. The latter paper presented from an ELF perspective some of the ideas contained in a slightly earlier work, Shohamy (2011). In the article, she points out that multilingual competence involves languages "bleeding" into each other, and thus differing in crucial ways from monolingual competence. And yet, she continues,

> this multilingual functioning receives no attention in language testing practices. Further, multilingual users who rarely reach language proficiency in each of the languages that is identical to that of their monolingual counterparts are always being compared to them and thus receive lower scores. Consequently, they are penalized for their multilingual competencies, sending a message that multilingual knowledge is a liability (p. 418).

These kinds of arguments linking English language testing with multilingualism are likely to have a major influence on ELF researchers' approaches to assessing ELF. This is particularly so in relation to what Jenkins (2015) has elsewhere described as "ELF 3," that is, English seen as a *multi*lingua franca, in which ELF's multilingual nature is its primary characteristic, rather than one feature among several.

This takes us directly to the second body of scholarship mentioned earlier, that is, multilingualism research. Key areas here that are likely to influence thinking about assessing ELF are the notion of dynamic bilingualism (Flores and Schissel 2014;

García 2009), the "multilingual turn" (e.g., May 2014), and translanguaging, that is, "fluid practices that go *between* and *beyond* socially constructed language and educational systems, structures and practices to engage diverse students' multiple meaning-making systems and subjectivities" (García and Wei 2014, p. 3; their italics). In all these, as well as in several other approaches including the work of scholars such as Cenoz (e.g., 2009) and Kirkpatrick (e.g., 2007), who have long argued that monolinguals should not provide benchmarks for the assessment of multilinguals' English, multilingualism is seen as the norm, monolingualism as the exception, and translanguaging as part of normal language practices. This work has profound implications for the future testing of English, and we will return to it in our discussion of future directions below.

We turn now to major contributions to assessing ELF from within ELF research itself. Despite the burgeoning amount of research into ELF that has been published over recent years, remarkably little has in fact focused on the issue of assessment. This is not surprising given the massive challenge that ELF presents for assessment. Nevertheless, five key publications have engaged with this challenge, all arguing in various ways that English language assessment urgently needs to address the findings of ELF research. Meanwhile, another group of publications has explored ELF in important ways in respect of the CEFR more specifically. We will now look at both groups.

Two of these publications (Chopin 2014 and Newbold 2014) were published in the same edited volume on ELF pedagogy. Of the 14 chapters in the volume, these were the only two focusing on assessment, demonstrating the current difficulties facing anyone attempting to conceive of ways of assessing English use that are not predicated on some kind of linguistic norms. Both authors focus on testing in higher education, and specifically on tests for university entrance. This is not surprising given that the rapid increase in mobility affecting higher education, with the result that many universities are becoming major sites of ELF communication. Both authors argue for the need to move away from native English norms in assessing suitability for university study. Chopin proposes a radical solution to the issue of language form: "simply to ignore it" (p. 200). She explains as follows:

> That is to say that language testing could and should change focus, away from form and towards other aspects of performance which may be more meaningful in terms of how people successfully communicate with each other (ibid.).

Chopin goes on to argue that "[f]orm, if the test-taker's speech is intelligible, could be in this way side-stepped" and that both native and nonnative speakers would need to be tested as both "would need to show evidence of being able to accommodate and negotiate meaning with interlocutors" (p. 201). She adds that "the native speaker would no longer be given a free pass, with the assumption that being a native speaker by definition gives an ability to communicate effectively in ELF settings" (ibid.).

In his chapter, Newbold describes an online test, Test of English for European University Students, that he and colleagues at the University of Venice devised to

test incoming Erasmus students' receptive skills. The listening test, which was most realistic in terms of ELF communication, or "the closest encounter with ELF" as Newbold himself puts it (p. 217), included a section in which candidates listened to students from a wide range of L1s exchanging opinions after a lecture. The candidates' response was overwhelmingly favorable with, for example, 47% saying it was "fairly realistic" and 53% "very realistic." The greater challenge, however, as Newbold points out, will be developing an ELF-oriented test of productive skills. Taking a very similar position to that of Chopin, he considers that any such test

> would need to be grounded in the pragmatics of ELF interaction, and it would need to identify features of successful communication, and to allow for formal variation in a qualitatively different way from rating scales currently used in institutionalized testing. It would need to be user-centred and norm-defocused (p. 220).

Finally, he argues, anyone devising such a test would need to have a clear understanding of its purpose and therefore of the precise context in which it was to be administered. This echoes Blommaert's point (see above).

Hall (2014) shares very similar concerns to those of the previous two authors. He challenges the monolithic approach of current standardized English language tests, with their focus on standard English and native English (which amount to the same thing for the examination boards) on the grounds that they ignore the diversity of English language users and uses. He argues "on both cognitive and social grounds... that the Englishes encountered and appropriated by non-native speakers will inevitably be qualitatively different from 'standard English' models, and that the effectiveness of the resources learners do develop should be assessed, where appropriate, independently of linguistic criteria" (p. 376). Hall proposes instead an approach that he calls "Englishing," a shift "from testing how people use the language to testing what they can *do* with it" (ibid.: his italics). This, he notes, would facilitate a move away from the problematic assumption of "a fixed notion of what English is, such that 'it' can be used" (p. 384), and will in his view will mean not merely making adjustments to current practices, but will involve a fundamental rethink of testers as to "what English is and how it is learned and used" (ibid.).

Jenkins and Leung (2014) take a similarly radical view of the need for change. After reviewing a range of so-called international English language tests, they argue from an ELF standpoint that these tests are far from international as they "continue to focus narrowly on native English norms, while no substantial adjustments have been made to the basic assumptions of what English is," with the result that they have "failed to keep in touch with contemporary developments in English" (p. 1613). They argue, citing McNamara (2011) that current English language testing ideology is having a negative impact on the language itself as well as on test candidates and their future prospects, and reaffirm the call to language researchers to contribute to the task of better understanding "what communication may comprise in terms of participant-driven uses of English as a linguistic resource in contemporary conditions" (p. 1615; also see Jenkins (2006) for an earlier discussion on this point).

Finally, focusing on English language university entry tests specifically, Jenkins (2016) takes up the previous theme by arguing that awareness of the sociolinguistic implications of the international spread of English, and of relevant findings in ELF research, is lacking in current test design. She points out that standardized tests are unable to cope with the fact that language is messy, and lingua franca use is even messier, which renders futile the attempt to impose a preset template on contingent use in diverse English contexts. To this extent, she argues, none of the current "international" examinations are fit for purpose. She concludes that to be valid, an international English language entry test

> will focus on everyone's ability to use English as a tool of intercultural communication in their own context, *not* on NNESs' ability to mimic certain anonymous NESs. And it will not allow NESs to see themselves as English language experts. This will give English language entry tests authenticity and validity, whereas currently they have neither as far as my international student research participants were concerned

The remaining contributions from ELF researchers focus specifically on the CEFR rather than on tests. It is important to consider the CEFR as it is currently used as the benchmark for so many tests in countries all around the globe, despite its original European basis. And in most of these countries, i.e., in the expanding circle, the test candidates will use their English in ELF communication rather than with speakers of the kind of native English on which they are tested. The problem with the CEFR in this respect is that it does not distinguish between a foreign language and a lingua franca. Furthermore, it tacitly assumes that language learning is about learning the native variety. Therefore when it is used as a reference framework for assessment, test development slipstreams into adopting this assumption. And because of its seemingly all-purpose supranational status and its global reach, it has pervasive influence that is difficult to discount. This has been accentuated by its remoteness of origin (hailing from a quasi-governmental organization in Europe), which has also made change difficult. Finally, its focus on target language competence has not been augmented, to date, by discussion on the learning processes that takes dynamic multilingualism and translanguaging into account.

As mentioned earlier, McNamara (2011, 2012) has drawn on an ELF perspective to critique the CEFR. He questions the source of the CEFR's authority and argues "the determination of test constructs [such as the CEFR] within policy-related frameworks leads to inflexibility" (2011, p. 500). Hynninen (2014) takes a similarly critical view of the CEFR from an ELF perspective. She considers that "the native speaker and native language culture foci particularly of the proficiency level descriptors" require a fundamental rethink in light of "the ways ELF speakers have been found to regulate language in ELF interactions" (p. 293). This, she argues, means "not only moving beyond native speaker-non native speaker contacts and the idea of a native speaker target culture, but also moving towards more context-aware assessment criteria" informed by ELF research findings (ibid.). Pitzl (2015) challenges the CEFR still further, and from a different standpoint: its representation and discursive construction of misunderstanding and communication breakdown. She provides

closely analyzed evidence to demonstrate that "the deficit portrait of intercultural communication in the CEFR may be based on a number of implicit logical fallacies, such as the idealized notion that L1 communication is perfect and devoid of miscommunication" (p. 91). She argues that essentialist notions of this kind should be abandoned in favor of conceiving "*understanding* as a jointly negotiated and interactional process" (ibid.; her italics), as proposed by researchers into Business ELF and intercultural communication.

Despite this small but growing body of focusing on assessment from an ELF perspective, now dating back 10 years, as we mentioned earlier, there has so far been little action in addressing this research on the part of the international examination boards, a situation that Newbold rightly describes as "somewhat surprising" (2014, p. 219).

## Work in Progress

It is now clear that the conceptual basis for a shift in thinking is in place for embracing the relevance of ELF research in language assessment. In the next period, research would need to pay attention to, *inter alia*, the following issues:

In psychometrically oriented standardised testing, construct validity has been a key criterion for test quality. Construct has been defined by Bachman and Palmer (1996, p. 89) as ". . . the precise nature of the ability we want to measure, by defining it abstractly." They further suggest that ". . . we can consider a construct to be the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task" (Bachman and Palmer 2010, p. 43). For the sake of argument, let us assume that we define ELF competence abstractly as "the ability to convey and negotiate meaning to achieve communicative goal/s." This abstract definition would not get us very far the moment we try to be specific about any assessment task because, as we have seen, the essence of ELF-mediated interaction is its agentive, contingent, dynamic and emergent nature. It is not possible to pre-specify the way/s in which communication is accomplished. The question here is whether the notion of construct, as it has been understood in language assessment research hitherto, is workable in assessment settings where ELF is a legitimate concern. (For a wider discussion on further work on "construct," see Newton and Shaw 2014).

In conventional rating scales, competence is generally graded in terms of levels, bands, stages, or marks, and the scoring rubrics tend to refer to performance descriptors that indicate the "typical" performance at different levels of accomplishment. All of this presupposes that we have a clear view of what different levels of accomplishment look like. Given the contingent nature of ELF-mediated communication, it would be difficult, indeed meaningless, to prespecify different levels. If this is the case, do we need to adopt a binary rating frame that comprises only "pass" or "fail"? If we did adopt this binary rating approach, would we be dealing with an exclusively "communication outcome" orientation in language assessment? Would we be jettisoning any possible use of assessment for formative language learning

purposes (because assessment activities would now be insensitive to different ways of using language and extents of accomplishment)? Or is it the case that ELF-minded assessment would need to engage in further empirical work to establish patterns of flexible use of linguistic and other semiotic resources so that we can begin to establish helpful vignettes of degrees and/or types of successful communication as reference points for assessment schemes?

All assessment of second/foreign/additional language implicates speakers' bi/multilingualism experience, particularly in terms of their learning experience of which lingua franca communication (in English or any other language) is likely to be a component. Does this mean that ELF considerations are automatically relevant to all language assessment, or is there a need to develop a finer-grained understanding of the relevance of ELF in relation to assessment purpose? It would be relatively easy to see the relevance of ELF sensibilities if we are considering the spoken English language competence of, say, engineers from diverse language backgrounds working in multinational and multilingual teams. ELF sensibilities might not be so obviously relevant if we are dealing with the criteria for assessing English language competence of, say, legal professionals who need to have very high levels of lexicogrammatical accuracy and idiomatic control in the written mode in accordance with a particular local jurisdiction. A good deal of conceptual and empirical work is needed to pave the way for better understanding of this complex question.

## Problems and Difficulties

It will be clear from all we have said so far, that the notion of assessing ELF involves a range of problems and difficulties, both practical and conceptual. At the practical level, the complexity of ELF communication – its emergent nature and the sheer diversity of its users – means that conventional testing methods involving testing against certain established norms cannot be used. Some of the ELF scholars cited above have suggested alternative areas for ELF assessment to focus on, such as "Englishing" (Hall 2014), which seem promising. But these alternatives will involve a major rethink and investment on the part of the examination boards, who seem, at least for now, not to have changed their practices, given that these practices are currently proving highly commercially successful. Until these frameworks are revised to incorporate lingua francas, and ELF in particular, English language assessment in general will continue to benchmark their criteria against an often inappropriate monolingual version of English.

A related issue is the powerful influence of established transnational and national English language assessment frameworks in different world locations. These frameworks often achieve their preeminence through complex processes of ideological articulation and political endorsements. There is still a tendency by the "big tests" to treat some distinctly Anglo-centric practices as "international." Some assessment frameworks such as IELTS present themselves as international because they are marketed internationally. The assumptions and norms underpinning their operations

are palpably based on Anglophone practices. Any attempt at reform or further development by language assessment professionals is likely to be a very complex and long-term effort (for a wider discussion on global trends in assessment and evaluation see Meyer and Benavot 2013).

## Future Directions

Looking ahead, we believe that it will be important to widen our notion of "English" beyond the so-called standard varieties and narrowly defined norms, to include divergent, local, agentive ELF use comprising contingently and jointly agreed wordings, and sociopragmatics influenced by elements of speakers' dynamic multilingual and translingual communication where appropriate and necessary. This means that the hitherto strong emphasis on reproduction of monolingual English native speaker norms and practices should be applied only where such criterial considerations can be justified in terms of context and purpose. English Language assessment in different world locations should pay close attention to the ways in which English is used for different purposes in different kinds of multilingual settings. Practically, then, the design and development of assessment criteria, procedures, and tasks should take full account of local practices and embrace a variety of assessment formats, activities, and reporting instruments that can help sample and reflect learner/user performance adequately. In other words, we are talking not just about "assessing ELF" as such but about taking account of ELF use where appropriate in the conceptualization and design of English language assessment.

## Cross-References


- ▶ Assessing Multilingual Competence
- ▶ Critical Language Testing
- ▶ The Common European Framework of Reference (CEFR)


## Related Articles in the Encyclopedia of Language and Education


Ofelia García, Li Wei: From Researching Translanguaging to Translanguaging Research. In Volume: Research Methods in Language and Education

Ingrid Gogolin, Joanna Duarte: Superdiversity, Multilingualism, and Awareness. In Volume: Language Awareness and Multilingualism

Barbara Seidlhofer: English as Lingua Franca and Multilingualism. In Volume: Language Awareness and Multilingualism

Oleg Tarnopolsky: Nonnative Speaking Teachers of English as a Foreign Language. In Volume: Second and Foreign Language Education

## References


Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Blommaert, J. (2015). *Dialogues with ethnography: Notes on classics and how I read them (working paper)*. Tilburg: Baylon Center, Tilburg University.

Canagarajah, A. S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language? *Language Assessment Quarterly, 3*(3), 229–242.

Cenoz, J. (2009). *Towards multilingual education*. Bristol: Multilingual Matters.

Chopin, K. (2014). Reconceptualizing norms for language testing: Assessing Engish language proficiency from within an ELF framework. In Y. Bayyurt & A. Sumru (Eds.), *Current perspectives on pedagogy for English as a lingua franca*. Berlin: De Gruyter Mouton.

Davidson, F. (2006). World Englishes and test construction. In B. B. Kachru, Y. Kachru, & C. L. Nelson (Eds.), *The handbook of World Englishes* (pp. 709–717). Malden: Blackwell Publishing.

Davies, A., Hamp-Lyons, E., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22*(4), 571–584.

Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics, 26*, 282–301.

Flores, N., & Schissel, J. L. (2014). Dynamic bilingualism as the norm: Envisioning a heteroglossic approach to standards-based reform. *TESOL Quarterly, 48*(3), 454–479.

García, O. (2009). *Bilingual education in the 21st century.* Malden: Wiley-Blackwell.

García, O., & Wei, L. (2014). *Translanguaging. Language, bilingualism and education*. Houndmills, Basingstoke: Palgrave Macmillan.

Hall, C. (2014). Moving beyond tests of English to tests of 'Englishing'. *ELT Journal, 48*(4), 376–385.

Hynninen, N. (2014). The common European framework of reference from the perspective of English as a lingua franca: What can we learn from a focus on language regulation? *Journal of English as a Lingua Franca, 3*(2), 293–316.

Jenkins, J. (2006). The spread of EIL: a testing time for testers. *English Language Teaching Journal 40*(1), 42–50.

Jenkins, J. (2015). Repositioning English and multilingualism within English as a Lingua Franca. *Englishes in Practice* (Working papers of the Centre for Global Englishes). De Gruyter Open: University of Southampton.

Jenkins, J. (2016). International tests of English: Are they fit for purpose? In H. Liao (Ed.), *Critical reflections on foreign language education: Globalization and local interventions*. Taipei: The Language Training and Testing Center (in press).

Jenkins, J., & Leung, C. (2014). English as a lingua franca. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. 4, pp. 1607–1616). Malden: Wiley.

Kachru, B. B. (1992). *The other tongue. English across cultures*. Champaign: University of Illinois Press.

Kachru, B. B. (2005). *Asian Englishes. Beyond the Canon*. Hong Kong: University of Hong Kong Press.

Kirkpatrick, A. (2007). Setting attainable and appropriate English language targets in multilingual settings: A case for Hong Kong. *International Journal of Applied Linguistics, 17*(3), 353–368.

Leung, C., & Street, B. (2014). Classroom constructions of language and literacy activity. In M. Prinsloo & C. Stroud (Eds.), *Educating for language and literacy diversity: Mobile selves* (pp. 23–44). Basingstoke: Palgrave Macmillan.

Lowenberg, P. (1993). Issues of validity in tests of English as a world language. *World Englishes, 12*(1), 95–106.

Lowenberg, P. (2000). Non-native varieties and the sociopolitics of English proficiency assessment. In J. Kelly Hall & W. G. Eggington (Eds.), *The sociopolitics of English language teaching* (pp. 67–85). Clevedon: Multilingual Matters.

Lowenberg, P. (2002). Assessing English proficiency in the expanding circle. *World Englishes, 21*(3), 431–435.

May, S. (Ed.). (2014). *The multilingual turn. Implications for SLA, TESOL and bilingual education*. New York/London: Routledge.

McNamara, T. (2011). Managing learning: Authority and language assessment. *LanguageTeaching, 44*(4), 500–515.

McNamara, T. (2012). English as a lingua franca: The challenge for language testing. *Journal of English as a Lingua Franca, 1*(1), 199–202.

McNamara, T. (2014). Thirty years on – Evolution or revolution. *Language Assessment Quarterly, 11*, 226–232.

Meyer, H.-D., & Benavot, A. (2013). PISA and the globalization of education governance: Some puzzles and problems. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governanace* (pp. 7–26). Oxford: Symposium Books.

Newbold, D. (2014). Engaging with ELF in an entrance test for European university students. In Y. Bayyurt & A. Sumru (Eds.), *Current perspectives on pedagogy for English as a lingua franca*. Berlin: De Gruyter Mouton.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.

Pitzl, M.-L. (2015). Understanding and misunderstanding in the common European framework of reference: What can we learn from research on BELF and intercultural communication? *Journal of English as a Lingua Franca, 4*(1), 91–124.

Seidlhofer, B. (2011). *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman/Pearson Education.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. London/New York: Routledge.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal, 95*(3), 418–429.

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal, 60*(1), 51–60.

# Part II

# Methods of Assessment

# Task and Performance-Based Assessment

Gillian Wigglesworth and Kellie Frost

**Abstract**

The increasing importance of performance testing in testing and assessment contexts has meant that the behavior of test tasks, how they perform, and how they are assessed has become a considerable focus of research. During the 1990s, performance assessment evolved alongside the multicomponential models of language that were emerging, while, at the same time, detailed frameworks of task characteristics were discussed which provided basis for both test design and test-related research. In second-language acquisition research, tasks have long been an important focus of research although the focus has been different in the testing context where the impact of the properties and characteristics of tasks and how they impact on test scores has been explored, as has the role of raters in the process.

Recently, interests have moved beyond assessing the individual components of language proficiency – speaking, writing, reading, and listening – to include integrated tasks which add a further element of complexity to the assessment process by incorporating more than one skill, for example, reading a passage and completing a writing task based on this. These types of tasks contribute to the increasing authenticity of the assessment for real-life situations but because these types of tasks involve engaging skills and strategies that are not normally included in language testing, further elements of complexity are added. These are currently being addressed through a variety of research studies.

G. Wigglesworth (✉)
Research Unit for Indigenous Language, ARC Centre of Excellence for the Dynamics of Language, Faculty of Arts, University of Melbourne, Parkville, VIC, Australia
e-mail: g.wigglesworth@unimelb.edu.au; gillianw@unimelb.edu.au

K. Frost
Language Testing Research Centre, School of Languages and Linguistics, University of Melbourne, Parkville, VIC, Australia
e-mail: kmfrost@unimelb.edu.au

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_8

**Keywords**
Task-based performance assessment • Authenticity • Task difficulty • Speaking • Writing

## Contents



## Introduction

A performance test is "a test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed" (Davies et al. 1999, p. 144). In the assessment of second languages, tasks are designed to measure learners' productive language skills through performances which allow candidates to demonstrate the kinds of language skills that may be required in a real-world context. For example, a test candidate whose language is being evaluated for the purposes of entry into an English-speaking university or college might be asked to write a short academic essay, or an overseas-qualified doctor might participate in a job-specific role play with a "patient" interviewer. These kinds of assessments are increasingly used in specific workplace language evaluations and in educational contexts to evaluate language gains during a period of teaching.

The relationship between task and performance testing is a complex one. In the context of language testing and assessment, performance assessment has become increasingly important over the last three decades and has been the focus of substantial empirical investigation. Performance-based assessments can be more or less specific in terms of the language skills they are designed to assess. Tests such as the IELTS or TOEFL are large-scale, high-stakes tests which are designed to evaluate largely academic language skills, while others have proved a valuable tool for assessing candidate performance in specific vocational contexts (e.g., the Occupational English Test, which is used for assessing the language skills of overseas-trained medical professionals prior to accreditation in Australia).

The role of tasks in performance-based assessments has recently attracted considerable attention, both from a theoretical and a practical perspective. Generally, there is little agreement about where "task-based language assessment" sits in relation to language testing more generally; Bachman (2002) uses the term "task-based language performance assessment" (TBLPA), while others (e.g., Norris 2002; Mislevy et al. 2002) refer more generally to task-based language assessment, or

TBLA. However, Brown et al. (2002) define task-based language testing as a subset of performance-based language testing, clearly distinguishing between performance-based testing, in which tasks are merely a vehicle for eliciting language samples for rating, and task-based performance assessments in which tasks are used to elicit language to reflect the kind of real-world activities learners will be expected to perform and in which the focus is on interpreting the learners' abilities to use language to perform such tasks in the real world.

## Early Developments

Performance assessments have been used for the evaluation of second languages for at least half a century. McNamara (1996) argues that their development has been the result of two factors. The first stemmed from the need to evaluate the language of second-language learners entering English-speaking universities and from the need to ascertain the language abilities of second-language learners entering specific workplace contexts (e.g., doctors, nurses, flight controllers, pilots, teachers, tour guides). The second has resulted from the increasing focus in second-language learning and teaching on communicative language ability with its focus on the ability to use language communicatively and appropriately in different contexts. Bachman's (1990) model of language proficiency, further developed in Bachman and Palmer (1996), with its focus on the learners' abilities to use language has been hugely influential in developing the agenda for research into task and performance-based language assessments. For test candidates, this trend toward task and performance-based assessment means that they are evaluated on a much greater range of language skills than those traditionally measured by the more discrete, paper-and-pen-based tests. Thus, second-language task and performance assessments have evolved in parallel with increasingly multicomponential models of language ability. More communicative approaches to language learning and teaching have been necessitated by the need to assess language in use, rather than language as object. Building on Bachman's (1990) model of language ability, Bachman and Palmer (1996) articulate a detailed framework of task characteristics intended as the basis for both test design and test-related research. These characteristics focus on the setting, the test rubrics, the input to the task (both in terms of format and language input), the expected response (again in terms of format and language), and the relationship between the input and the response.

Second-language performance assessments can be conducted in a variety of contexts. One option is in situ (e.g., in the classroom, in the workplace) through observation. McNamara (1996, following Slater, 1980 and Jones, 1985) calls this a "direct assessment" since the language behavior is being evaluated in the context in which it is being used. Alternatively, second-language performance assessments may be evaluated through simulations of real-world performance, i.e., tasks tailor-made for the particular communicative purpose of the assessment. McNamara (1996) argues that there are two factors which distinguish second-language performance tests from traditional tests of the second language: the fact that there is a

performance by the candidate and that this is judged using an agreed set of criteria. Norris et al. (1998) add a third criterion arguing that the tasks used in performance assessments should be as authentic as possible.

McNamara (1996) argues a distinction between *strong* and *weak* forms of second-language performance assessment, based on the criteria used for judging the performance. In the "strong" sense, assessment is made on the basis of the extent to which the actual task itself has been achieved, with language being the means for fulfilling the task requirements rather than an end in itself. In the "weak" sense, the focus of the assessment is less on the task and more on the language produced by the candidate, with the task serving only as the medium through which the language is elicited – successful performance of the task itself is not the focus of the assessment. This distinction is revisited in the later work of Brown et al. (2002, pp. 9–11) in which the term *performance-based testing* was used where the tasks are used to elicit language samples for the purposes of rating – in McNamara's terms, "weak" performance assessments – and *task-based performance assessments* involve assessments in which tasks are used to elicit language to reflect the kind of real-world activities learners will be expected to perform and in which the focus is on interpreting the learner's ability to perform such tasks in the real world (p. 11), "strong" performance assessments in McNamara's terminology. This provides two very different ways of defining the construct. In the "weak" version, the construct is defined as language ability. In the "strong" version, it includes everything which might contribute to the successful completion of the task, which means that there are more likely to be a range of confounding factors including task characteristics and test taker interactions with these that might affect score interpretation and use.

## Major Contributions

In the second-language acquisition (SLA) literature, the properties and characteristics of tasks, and the different conditions under which they can be administered, have been the subject of intense scrutiny. A major focus of this research has been on how learners manage the differential cognitive load associated with different types of tasks and the extent to which these varying conditions and characteristics influence learner productions (see, e.g., Foster and Skehan 1996; Skehan and Foster 1997; Ellis 2003; Yuan and Ellis 2003; Ellis and Yuan 2004; Robinson 2007; Tavakoli and Foster 2008). Different variables have been systematically investigated incorporating the conditions under which the tasks are administered, i.e., those conditions external to the task. The task condition which has received considerable attention is the provision, or not, of varying amounts of planning time (see, e.g., Ellis 2005). The internal characteristics of tasks have also attracted substantial attention. In particular, the series of studies by Foster and Skehan (1996, 1999) and Skehan and Foster (1997, 1999) indicate that different task characteristics (e.g., dialogic versus monologic, structured versus unstructured, simple versus complex in outcome) have differential impacts on measures of fluency, complexity, and accuracy in the learners' discourse (Skehan 2001). Much of the above work has been motivated by

information-processing models of second-language acquisition (see Skehan 1998) and has used detailed analyses of elicited discourse (written or spoken) to evaluate changes in measures of complexity, accuracy, and fluency which might result from different task conditions and characteristics.

In relation to performance testing and assessment, the need to link test tasks to theoretical models of cognition and language learning is evident in Mislevy, Steinberg, and Almond's (2003) "evidence-centered" approach to designing assessments and in Kane's (2006) highly influential argument-based approach to test validation. Studies have focused on exploring how different task properties might impact on candidate performance in the context of classroom-based assessment practice and in relation to high-stakes assessments, such as TOEFL and IELTS. The approach taken by many of these studies has been to evaluate the learner performances on two levels – externally through rating and internally through analyses of candidate discourse.

Task-based performance assessments in teaching programs have proved particularly valuable because task-based assessments can be linked to teaching outcomes, provided outcomes are defined in terms of task fulfillment, rather than purely in terms of language ability. A further consequence can be that well-designed assessment tasks have the potential to provide positive washback into the classroom. However, the issues raised by the use of tasks for these types of assessments are considerable. Brindley and Slatyer (2002) examined the effect of varying the characteristics and conditions in listening assessment tasks used in the context of an outcome-based reporting system in which teachers themselves develop tasks for assessment purposes, and Wigglesworth (2001) undertook a similar investigation of speaking tasks by manipulating a series of task conditions and characteristics. Both studies found small effects as a result of manipulating the variables but also point out that interaction effects impact on the variables in ways which are difficult to separate. Such studies, which systematically manipulate different task variables, are of crucial importance since teachers are often involved in the development of assessment tasks and must understand how these work in order to produce comparable and defensible judgments of students for classroom assessment purposes.

In the high-stakes testing context, the impact of task properties and characteristics on performance has been investigated in a series of studies which used test scores to investigate potential differences (e.g., Lee 2006), as well as measures of complexity, accuracy, and fluency to determine whether finer distinctions imperceptible to raters are marked in the candidate discourse (see, e.g., Iwashita et al. 2001; Elder et al. 2002; Wigglesworth 1997; Brown et al. 2005; Elder and Wigglesworth 2005). The general outcome of these studies has been that raters perceive no differences, and in general, very few, if any, differences have been detected in the discourse. Necessarily, given the testing focus, task difficulty has been a particular focus of these studies, since for testing purposes, it would be useful to be able to design tasks of predictable levels of difficulty which can be manipulated to elicit appropriate performances across candidates. Norris et al. (1998) and Brown et al. (2002) provide a comprehensive empirical investigation of the problems of the comparability of real-world performance tasks, by systematically manipulating three cognitive processing

variables (code complexity, cognitive complexity, and communicative demand) in a series of test tasks. In summarizing their findings in relation to task difficulty, Norris et al. (2002, p. 414) point out the importance of individual responses to tasks, which may impact on measures of task difficulty. They argue that:

> initial evidence from this study did not support the use of the cognitive processing factors – as operationalized in our original task difficulty framework – for the estimation of eventual performance difficulty differences among test tasks. While there was some indication that average performance levels associated with the three cognitive task types differed in predicted ways, these differences did not extend to individual tasks. What is more, evidence suggests that examinees may have been responding to tasks in idiosyncratic ways, in particular as a result of their familiarity with both task content and task procedures.

Elder et al. (2002) asked candidates about their perception of task difficulty and found they too were unable to estimate the difficulty of a task even after they had performed it. Indeed, Bachman (2002) argues that the complex nature of task performances, which involve large numbers of interactions (e.g., between candidate and task, task and rater, candidate and interlocutor, etc.), means that task difficulty cannot be conceptualized as a separate factor. Specifically, in relation to speaking tests, Fulcher and Reiter (2003) question assumptions that underlie SLA approaches to conceptualizing task difficulty in terms of particular task conditions and characteristics, suggesting instead that difficulty is more likely explained by interactions between the pragmatic features of tasks and the first-language background of test takers.

While both writing and speaking performance test tasks need to be subjectively rated, with all that rater variables entail, performance testing in the assessment of speaking skills brings the additional variable of the interlocutor. As Brown (2003) shows, the same candidate can produce qualitatively different performances when interviewed by different interviewers, and this may mean that the raters interpret the candidate's performance differently. Other studies (e.g., Morton et al. 1997; McNamara and Lumley 1997; Davis 2009; May 2009), where raters evaluated not only the candidate but the interlocutor performance as well, have found that raters tend to compensate for what they view as deficient interviewer behavior. Studies by Ducasse and Brown (2009) and Galaczi (2014) suggest that interactional features beyond topic development and organization, such as listener support strategies or interactional listening, turn-taking behaviors, and interactional management, should be included in rating scales.

Another aspect of a task which may influence the test scores is the nature of the rating scale used to judge performance. Since these judgments are by nature subjective, they require well-defined rating scales. Rating scales consist of a set of criteria upon which a performance can be judged. They are necessarily limited in scope because no rating scale can attend to all possible aspects of performance, and thus choices about *what* to rate (intelligibility, accuracy, complexity, clarity) must be made, as well as choices about what *proportion* of the score is appropriate to allocate to each rating criterion – in other words, some criteria may be weighted more heavily than others. Rating scales need to be designed to allow accurate judgments of the

speech or writing samples elicited and need to be valid in terms of the relevant language construct. Rating scales may rate task performance globally, based on a holistic impression, or analytically on a feature-by-feature basis. Knoch (2009) compared two rating scales, holistic (consisting of general descriptors) and analytic, consisting of detailed, empirically derived descriptors. She found that the latter scale was associated with higher rater reliability and was preferred by raters. Fulcher et al. (2011) distinguish between two broad approaches to rating scale design and development: measurement-driven approaches, whereby descriptors are ordered in a linear fashion on a single scale, and performance data-driven approaches, whereby descriptors are empirically derived. The researchers argue that the latter approach provides richer and more meaningful descriptions of performances.

Rating scales can only ever guide human judgments, however, and decisions between raters may vary widely, with potential consequences for test fairness. It is now widely acknowledged that raters differ in both self-consistency and in their severity (Upshur and Turner 1999; Huhta et al. 2014; Granfeldt and Malin 2014) and also in the way they construe the different elements of the rating scale (Lumley 2002; Harding et al. 2011; Kuiken and Vedder 2014). Rater training thus becomes a critical component in task-based performance assessment. While ideally rater training may aim to reduce differences in severity across different raters, where this is not achievable, training needs to ensure that raters discriminate consistently in terms of severity across different levels of performance. As a result of these inherent differences in rater severity, best practice in assessment advocates double rating or even multiple ratings in the event of discrepancy between pairs. Statistical analyses of scores can then be used to gain a greater understanding of how different raters behave or to compensate for individual rater differences.

## Work in Progress

A central tenet of task-based language assessments is that the tasks are designed to represent authentic activities which test candidates might be expected to encounter in the real world outside the classroom. In particular, as Douglas (2000) points out, authenticity is central to the assessment of language for specific purposes and is part of what differentiates it from more general types of language testing. This is because a "specific purpose language test is one in which test content and methods are derived from an analysis of a specific purposes target language use situation, so that test tasks and content are authentically representative of tasks in the target situation" (p. 19). However, the issue of authenticity is not a trivial one, and the extent to which specific tasks can represent authentic real-world activity has attracted considerable debate and empirical investigation, using a variety of different approaches (see, e.g., Cumming et al. 2004; Lewkowicz 2000; Spence-Brown 2001; Wu and Stansfield 2001).

While performance-based tests have traditionally focused on independently measuring the four core language skills (speaking, writing, listening, and reading), efforts to better simulate real-world task demands, thereby enhancing authenticity,

have led to the development and use of integrated speaking and writing tasks (e.g., the TOEFL Internet-based test (iBT)). Integrated tasks require test takers to read or listen to source texts and to incorporate information from these texts into their speaking or writing test performances (Lewkowicz 1997). In addition to enhancing the authenticity of the tasks, integrated tasks also mitigate against some candidates having greater familiarity with the topic than others, since a common source of input is provided.

Existing research into the use of integrated writing tasks has examined how writers make use of the source material when responding to integrated tasks (e.g., Cumming et al. 2006; Plakans 2009; Weigle and Parker 2012), as well as the discourse produced by students across different score levels on the writing section of the TOEFL iBT (Gebril and Plakans 2013; Plakans and Gebril 2013). Studies addressing the use of integrated tasks as a measure of speaking ability have examined test takers' strategic behaviors (Barkaoui et al. 2013), rater orientations to integrated tasks (Brown et al. 2005), the impact of task type on test scores (Lee 2006), and the way in which test takers incorporate source materials into spoken performances (Brown et al 2005; Frost et al. 2012). In a recent study, Crossley et al. (2014) examine the interaction between test takers' spoken discourse, characteristics of task and stimulus materials, and rater judgments of speaking proficiency on a listening-speaking task of the TOEFL iBT. They found that the integration of source text words into spoken performances was predicted by three-word properties: incidence of word occurrence in the source text, the use of words in positive connective clauses, and word frequency in the source text. They also found that the incidence of source text words in the spoken responses was a strong predictor of human judgments of speaking quality.

## Problems and Difficulties

While there is broad agreement that task authenticity is desirable in performance testing and assessment (e.g., Bachman and Palmer 1996; Douglas 2000; Norris et al. 1998; Brown et al 2002), the extent to which inferences can be made from the language elicited by particular test tasks as a reflection of the candidates' ability to manage the task in a subsequent real-world context is not fully resolved.

Concerns that need to be addressed in relation to authenticity relate to the problem of the generalizability of the outcome. In the "weak" view of language testing, where concern is with the underlying language abilities, a criterion of task fulfillment may not be considered of great importance. In the "strong" view of performance testing, a task designed to assess the ability of candidates to carry out the activity in a real-world setting would need to be assessed on a criterion of task fulfillment rather than for its linguistic accuracy, for example. An unresolved issue here is who should decide whether the task has been carried out successfully – language specialists or specialists in the field of the task activity? The gap between linguistic criteria and the aspects of communication valued by professionals in the workplace, for example, is widely acknowledged. There are a number of studies which have examined this issue

(e.g., Elder and Brown 1997; Brown 1995; Elder 1993; Elder et al. 2012; Knoch 2014; Kim and Elder 2015), but the question remains one of balancing authenticity and generalizability. While the "weak" view is likely to assess underlying language skills in ways which are relatively broadly generalizable, the "strong" view is likely to produce judgments which are more authentic and relevant to the real-life situations toward which the candidate may be moving. These judgments about the quality of performance may not, however, be replicable in other contexts.

Task-based performance testing is attractive as an assessment option because its goal is to elicit language samples which measure the breadth of linguistic ability in candidates and because it aims to elicit samples of communicative language (language in use) through tasks which replicate the kinds of activities which candidates are likely to encounter in the real world. As a test method, however, it remains one of the most expensive approaches to assessment and, in terms of development and delivery, one of the most complex. There is also the potential for reduced generalizability since tasks used in such assessments tend to be complex and context specific, which means that inferences which are based on them may not always extrapolate to the domains they are intended to represent. An additional difficulty is that of replicating tasks in a way which ensures consistency of measurement.

## Future Directions

The development of appropriate tasks for use in performance assessment must be underpinned by an understanding of how the tasks relate to the construct and of which factors may potentially interfere with their validity and reliability. There is currently only a relatively limited amount of empirical research which systematically examines the types of tasks used in task and performance-based assessments and which can illuminate how different tasks work for assessment purposes. The complex nature of tasks, and their relationship to real-world performances, makes it crucial that we understand more about how the various different elements of the task, which impact on candidate performance with the task, interact.

Performance on integrated tasks, for example, requires candidates to engage skills and strategies that may extend beyond language proficiency in ways that can be difficult to define and measure for testing purposes. As Douglas (1997) and Lee (2006) have noted, test taker performances on integrated tasks involve not only productive skills but also comprehension skills and the ways in which these dimensions of language ability are integrated by test takers into their language performances remains, as yet, predominantly intuited by test developers. Furthermore, while it is well known that stimulus materials impact on test performance, the way in which test takers make use of these materials in their responses, particularly the strategies involved in summarizing and incorporating content from written and oral texts into speaking performances, is not well understood and requires further empirical investigation.

Testing is a socially situated activity although the social aspects of testing have been relatively under-explored (but see McNamara and Roever 2006). Testing and

assessment activities take place in a social context, and this is particularly the case with task- and performance-based assessment. In speaking assessments, the interlocutor has a crucial role to play. However, while the interlocutor is often a trained interviewer, this role may also be taken by another test candidate or a group of test candidates. In relation to paired and group test activities, a whole raft of variables are ripe for exploration since "we can hypothesize that the sociocultural norms of interaction … contribute significantly to variability in performance" (O'Sullivan 2002, p. 291). The extent to which they contribute in systematic ways to the way tasks are interpreted and undertaken is yet to be determined.

## Cross-References


- ▶ Assessing Meaning
- ▶ Assessing Students' Content Knowledge and Language Proficiency
- ▶ Dynamic Assessment
- ▶ Language Assessment Literacy


## Related Articles in the Encyclopedia of Language and Education

Klaus Brandl: Task-Based Instruction and Teacher Training. In Volume: Second and Foreign Language Education

Martin East: Task-Based Teaching and Learning: Pedagogical Implications. In Volume: Second and Foreign Language Education

Marta Gonzales-Lloret: Technology and Task Based Language Teaching. In Volume: Language, Education and Technology

## References


Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453–476.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*, 304–324.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369–394.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1–15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25.

Brown, J. D., Hudson, T., Norris, J., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments* (Technical report, Vol. 24). Honolulu: University of Hawaii Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series, Vol. MS-29). Princeton: Educational Testing Service.

Crossley, S., Clevinger, A., & Kim, Y. J. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly, 11*(3), 250–270.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(2), 107–145.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series, Vol. MS-30). Princeton: Educational Testing Service.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series, Vol. MS-8). Princeton: Educational Testing Service.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443.

Elder, C. (1993). How do subject specialists construe classroom language proficiency. *Language Testing, 10*(3), 235–254.

Elder, C., & Brown, A. (1997). Performance testing for the professions: Language proficiency or strategic competence? *Melbourne Papers in Language Testing, 6*(1), 68–78.

Elder, C., & Wigglesworth, G. (2005). An investigation of the effectiveness and validity of planning time in part 2 of the oral module. Report for IELTS Australia.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347–368.

Elder, C., Pill, J., Woodword-Kron, R., McNamara, T., Manias, E., McColl, G., & Webb, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly, 46*(2), 409–419.

Ellis, R. (2003). *Task based language learning*. Oxford: Oxford University Press.

Ellis, R. (Ed.). (2005). *Planning and task performance in a second language*. Philadelphia: John Benjamins.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research, 3*, 299–324.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening speaking task: A discourse based analysis of test takers' oral performances. *Language Testing, 29*(3), 345–369.

Fulcher, G., & Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321–344.

Fulcher, D., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9–27.

Granfeldt, J., & Argen, M. (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing, 31*(3), 285–305.

Harding, L., Pill, J., & Ryan, K. (2011). Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly, 8*, 108–126.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307–328.

Iwashita, N., Elder, C., & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401–436.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York: Macmillan.

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing, 32*(2), 129–149.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Knoch, U. (2014). Using subject specialist to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes, 33*, 77–86.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*, 131–166.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of language and education* (Language Testing and assessment, Vol. 7, pp. 121–130). Dortrecht: Kluwer.

Lewkowicz, J. (2000). Authenticity in language testing. *Language Testing, 17*(1), 43–64.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397–421.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140–156.

McNamara, T., & Roever, C. (2006). *Language testing: The social turn.* London: Blackwell.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Morton, J., Wigglesworth, G., & Williams, D. (1997). Approaches to validation: Evaluating interviewer performance in oral interaction tests. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 175–196). Sydney: NCELTR.

Norris, J. (2002). Interpretations, intended uses and designed in task-based language assessment. *Language Testing, 19*(4), 337–346.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawaii Press.

Norris, J. M., Brown, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395–418.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*, 561–587.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(2), 217–230.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching, 45*(3), 193–213.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167–187). Longman: Harlow.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influence on foreign language performance. *Language Teaching Research, 1*, 185–211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retelling. *Language Learning, 49*(1), 93–120.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*(4), 463–481.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 58*(2), 439–473.

Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82–111.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing, 21*, 118–133.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85–106.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 186–209). Harlow: Longman.

Wu, W., & Stansfield, C. (2001). Toward authenticity of task in test development. *Language Testing, 18*(2), 187–206.

Yuan, F., & Ellis, R. (2003). The effects of pretask planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics, 24*, 1–27.

# Using Portfolios for Assessment/Alternative Assessment

Janna Fox

**Abstract**

Alternative assessment has often been cast in opposition to traditional testing (particularly, high-stakes, discrete-point, multiple-choice testing). However, some (e.g., Bailey, K. *Learning About Language Assessment: Dilemmas, Decisions, and Directions*. Pacific Grove: Heinle & Heinle, 1998) have argued that it is more accurate to regard such tests and testing practices as one end of a continuum of assessment possibilities or alternatives *in* assessment. Although there are many alternative assessment approaches (see, e.g., criterion-referenced observational checklists, reading response journals, learning logs, poster presentations), the literature reviewed for this chapter suggests that over the past few years, portfolio assessment has become the most pervasive and prominent alternative assessment approach. Although portfolios take different forms and serve different purposes, they share in common the ongoing selection and collection of work as evidence of learning and development over time. Portfolio assessment initiatives have become increasingly used in language teaching and learning contexts, and their potential benefits have been widely promoted (e.g., Little, Lang Test 22(3):321–336, 2005), particularly when they serve formative/learning purposes; they have been less successful in summative assessment contexts (e.g., Fox, Contact Spec Res Symp Issue, 40(2), 68–83, 2014). Discussions of alternative assessment (which are at times viewed as more *authentic*, because they are closer to and have more in common with classroom practices) have continued to prompt lively discussions of validity and reliability. Arguably, however, the most substantive changes in alternative assessment have occurred as a result of the widespread use of increasingly sophisticated technologies. For example, e-portfolios have emerged as an important assessment alternative, which can provide a more flexible, less cumbersome, and longer-term

J. Fox (✉)
School of Linguistics and Language Studies, Carleton University, Ottawa, ON, Canada
e-mail: janna.fox@carleton.ca; Janna_Fox@carleton.ca

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_9

record of a student's development or a program's performance. Future alternative assessment approaches will continue to be increasingly enhanced by technological innovation, but such digital records may also generate concerns.

**Keywords**

Alternative assessment • Portfolio assessment • e-Portfolios • Technology in assessment

## Contents



## Introduction

Alternative assessment is most often discussed as an *alternative* to standardized tests and testing practices, which result in scores and normed comparisons of individuals and groups (e.g., Maslovaty and Kuzi 2002). Hargreaves et al. (2002) note that unlike a standardized test, alternative assessment is "designed to foster powerful, productive learning for students themselves" (p. 70). In addition to portfolios, they list conferences, observational checklists, self- or peer- assessment, diaries, learning logs, poster presentations, and projects, as examples of alternative assessment approaches. However, portfolio assessment is arguably the most pervasive and influential example of an alternative assessment approach.

In their recent consideration of the history of portfolio assessment, Carlson and Albright (2012) discuss "the inability for traditional tests (i.e. multiple choice, competency tests, essay tests) to reflect the students' *real* or *authentic* . . . abilities" (p. 102). Hamp-Lyons and Condon (2000) argue that portfolio assessment is a response to "the need to measure more complex phenomena" (p. 3), and note that such "newer assessment methods generally referred to as *authentic* assessment or *alternative* assessment are predominantly performance assessments" (p. 18).

Task-based approaches to assessment are also often mentioned as a "feature of alternative assessment," as part of "the move to base assessment not on multiple-choice format tests, but on actual instances of use by learners" (McNamara 1997, p. 132). For example, "embedded assessment tasks" (Spence-Brown 2001, p. 466) integrate assessment within teaching tasks to enhance interactiveness and engagement. Such tasks are consistent with *dynamic assessment* approaches (e.g., Leung 2007) and the view that learning develops from and is embedded in social interactions

with others, e.g., people, texts, objects, and events. In dynamic assessment, examiners provide either predetermined or spontaneous assistance (i.e., *mediation*) for a learner engaging in an instructional activity, in order to assess the learner's performance.

Alternative assessment is also associated with *accommodation* for students who have special needs. For example, alternative assessment has been devised for bilingual, English language learners (ELLs) who are studying in English-medium contexts. Some alternative tests allow ELLs to take achievement tests in content areas with the use or support of their first languages. Proponents argue that "in a multicultural, multilingual society, assessment policies must seek excellence and equity simultaneously, or they will accomplish neither" (Lacelle-Peterson and Rivera 1994, p. 57). They view alternative assessment as a means of encouraging greater educational equity.

Given that assessment is value laden in that what is valued in an assessment defines what is worth knowing or doing, some suggest that alternative assessment is more ethical (Lynch and Shaw 2005), democratic (Shohamy 2001), or sensitive to varying social conditions (Hamp-Lyons and Condon 2000).

## Early Developments

Early discussions of alternative assessment were dominated by opposing views. Some suggested that alternatives *in* assessment better defined the relationship between standardized testing practices and alternatives. They argued that describing alternative assessment and testing as polar opposites missed "an important point: that there may be many increments between these poles, and that shades of gray are possible" (Bailey 1998, p. 207). Viewing assessment as a continuum, Brown and Hudson (1998) argued that at the classroom level, "language teachers have always done assessment in one form or another and these new procedures are just new developments in that long tradition" (p. 657).

However, others (e.g., Lynch and Shaw 2005) argued that alternative assessment was diametrically opposed to traditional testing practices, because it focused on and valued the unique, personal, situated, and individual, whereas testing focused on (and valued) the replicable, the generalizable, and the group. Whereas Brown and Hudson (1998) viewed tests, portfolios, or observations as options in the continuum of practices that comprise the assessment repertoire, Lynch and Shaw (2005) took the paradigmatic view that alternative assessment was rooted in a cultural, epistemological, and axiological perspective that is fundamentally different from that of traditional testing in (1) the requirements for reliability and arguments for validity, (2) the implicit nature of stakeholder relationships (e.g., tester and test taker versus teacher and learner), and (3) the emphasis on learning products or scores rather than learning as an ongoing process. They argued that traditional tests and testing culture promote test-only strategies, whereas alternative assessment is informed by an assessment culture, which draws on multiple sources of evidence, drawn over time, to support learning and decision-making.

Shohamy (1996) linked developments in language testing and the emergence of what she refers to as an alternative era in assessment to changing theoretical definitions of "what it means to know a language" (p. 143). She argued that eras in testing reflect changing definitions of the language construct and identified traditional testing as part of a "discrete-point era" which, as Hamp-Lyons and Condon (2000) point out, was "fed by the rationalist-empiricist 1930s and 1940s (the period of behaviourism in psychology)" (p. 18) and the concomitant rise of structural linguistics, psychometrics, and educational measurement. Not only did discrete-point tests faithfully represent the language constructs of the day, they were mathematically tractable, met statistical requirements, and provided reliable tools for efficient decision-making.

However, as theoretically and empirically driven conceptions of language evolved, new eras in language testing emerged, which Shohamy identifies as the integrative era, the communicative era, and the performance testing era. She suggests the move to an alternative era, recognizing that "there are different types of language knowledge and mastering one type is no guarantee for mastering another ... different instruments are capable of *seeing* different things" (p. 152). She argued that it is impossible for a single test to measure the complex phenomena of language as we currently understand it, and therefore there is a need for "multiple assessment procedures" (p. 152). Shohamy suggested that *complementary assessment* was a more precise label than alternative assessment for this assessment approach.

Shohamy's reinterpretation of alternative assessment as multiple or complementary assessment is in sync with others, who conclude that alternative assessment is best exemplified by portfolio approaches to assessment (e.g., Hamp-Lyons and Condon 2000).

Although there are many purposes and uses for portfolios, in general they are repositories of artifacts (e.g., reflections, works in progress, self- and peer- assessments, final products) assembled over time as evidence of development, learning, or capability. In the classroom context, they typically require learners to actively participate in the assessment process by selecting which of their performances will be evaluated, collaborate with other students and the teacher in identifying criteria for evaluation, and reflect on their learning over time and in relation to accumulated evidence.

Although there have been a number of attempts to use portfolio assessment as an alternative to traditional tests in large-scale, high-stakes, or summative contexts (e.g., Koretz et al. 1994), the use of portfolio assessment for *formative* purposes has had a long and well-documented history. Portfolios have also played a prominent role in writing and composition studies since the 1980s and corresponded to the shift from product-oriented pedagogies to process-oriented ones.

## Major Contributions

In examining the evolution of alternative assessment, the contributions of Peter Elbow and Pat Belanoff figure prominently. In 1986, they published seminal research on their experimentation with a pass/fail portfolio system as an alternative

to traditional tests and essays in a university-level writing class. They argued that their use of portfolio-based assessment shifted the focus of classroom conversation from products, grades, and scores, to feedback, reflection, revision, and collaboration. Their research triggered a wave of interest in portfolio use in language teaching settings (e.g., Hamp-Lyons and Condon 2000). From the 1980s, portfolios (or writing folders) became a pervasive artifact in writing classrooms. The writer's composing *process* became the focus of classroom activity, which was characterized by drafting, peer conferencing, and iterative and recursive revision. Evidence of the process (e.g., successive drafts, conference checklists, reflective logs) was collected for ongoing and final assessment. The emphasis on the writer and the writing process was a dramatic shift away from the traditional focus on the text or *product* (i.e., the accuracy and quality of a finished essay), which had characterized earlier pedagogical approaches to writing. Thus, Elbow and Belanoff (1986) provide an early example of alternative assessment in practice.

In 1995, Huerta-Macías argued for assessment approaches in language teaching that were both "non-intrusive to the classroom [and] authentic" (p. 9). She suggested conferences, observational checklists, personal journals, work samples, and anecdotal records were better alternatives than traditional testing. Not only would such alternative assessment activities engage students in ongoing and active learning, but they would also increase the trustworthiness and usefulness of assessment, because multiple sources of evidence could be accumulated over time to account for learning, development, and achievement. Others (Delandshere and Petrosky 1998; Valdés and Figueroa 1994) made similar arguments, pointing out that detailed narrative profiles were far more useful outcomes of assessment than one-off, decontextualized numerical scores.

Huerta-Macías (1995) equated the increased authenticity of alternative assessment with both reliability and validity: "Alternative assessments are in and of themselves valid, due to the direct nature of the assessment" (p. 10). In keeping with these notions, Moss (1994) argued for a reconceptualization of reliability based on a *hermeneutic approach*, which acknowledges the situated, unique, and varying contexts of assessment. She viewed arguments for validity as internal to the assessment process itself, and reliant upon dialogue and consensus reached among key stakeholders. She questioned the traditional notion that generalizing from single (or multiple) test performance(s) to a population of possible performances was the ultimate goal of assessment.

Her perspective was deeply rooted within an interpretive or constructivist tradition, which views language as socially constructed and situated in contexts of use – rather than as an underlying trait or ability, which remains stable across contexts. As Maslovaty and Kuzi (2002) put it, "alternative assessment is based on the principles of constructivism in that it rests on authentic inquiry tasks which give significance to learning and are relevant to the real world of the learner" (p. 200).

Alternative assessment approaches have typically been informed by sociocultural theory, as Gipps (1999) notes: "By combining interpretive and sociocultural perspectives, we can begin to cast new light on the relationship and power dynamics between pupil and teacher in the context of assessment" (p. 356). From a sociocultural perspective, knowledge is situated and learning is social, interactive, collaborative, and embedded in the local cultural life of the individual.

This perspective is not shared by traditional testing which requires test takers to act autonomously and in response to tests that are external, formalized, and hierarchical (Lynch and Shaw 2005). The increased use of alternative assessment approaches has occurred alongside the development of critical applied linguistics (Lynch 2001), which considers tests as mediating tools in the overt (and covert) exercise of power (Shohamy 2001). Critical theorists have argued that tests define what is valued, and only what is valued tends to have currency in the classroom with learners, teachers, and other stakeholders. From a critical perspective, alternative assessment created the potential for the sharing of power and for the valuing of the individual, because it allowed for a more collaborative, dialogic interchange between the assessor and the assessed.

The growth in alternative assessment approaches has occurred at a time when qualitative and mixed methods research has gained prominence within applied linguistics generally and within language assessment specifically (e.g., Cheng and Fox 2013). Some researchers have linked the growth of alternative assessment to the increased use of qualitative approaches in research (e.g., Leung and Mohan 2004). Whereas quantitative research has long been associated with tests, testing practices, and psychometrics, qualitative research is consistent with alternative assessment approaches. Qualitative researchers assume that all human activity is both situated and embedded in contexts. They collect, analyze, and report on data as rich and thick descriptions of activity in context. This is consistent with the focus of alternative assessment, and particularly portfolio assessment, as learners in collaboration with teachers and their peers collect evidence of their learning activity over time. Such evidence is unique to their classroom context, variable, individual, but also *rich and thick*, in that it provides multiple sources of evidence of their learning development.

More recently, Cheng and Fox (2013) report on the increasing prevalence of mixed methods approaches in language assessment research. Mixed methods research uses both quantitative and qualitative methods and merges findings from each strand in responding to the research questions that guide a study. The notion of a continuum of alternatives *in* assessment, which extends from tests and testing practices (Brown and Hudson 1998) to alternative assessment approaches, is consistent with the steady expansion of mixed methods research. Mixed methods research moves beyond paradigmatic polarity and allows researchers to address more and varied questions about the complex phenomena that are the focus of applied linguistics in general and language assessment in particular.

Increasing interest in alternative assessment approaches is also related to the growing interest in classroom assessment practices and the role of formative assessment, which has been richly re-theorized within both language learning contexts (e.g., Rea-Dickins 2001) and broader educational contexts (e.g., Black and Wiliam 2006).

## Work in Progress

There are a number of recent trends in alternative assessment which warrant particular discussion:

1. *The increasing use of portfolio assessment across purposes and contexts*

As noted above, over the past few years, portfolio approaches have become a prevalent feature in language teaching and learning and have also become a preferred assessment approach for monitoring and evaluating language learners' proficiency development and achievement. For example, in Europe, the English Language Portfolio (ELP) is used to support and document proficiency development in relation to the benchmark criteria provided by the Common European Framework of Reference (CEFR). In Canada, Portfolio-Based Language Assessment (PBLA) plays a similar role in relation to the Canadian Language Benchmarks (CLB). Correspondingly, in the United States, LinguaFolio and the Global Language Portfolio are referenced to the American Council on the Teaching of Foreign Languages (ACTFL) proficiency scale.

Proponents of these portfolio approaches suggest that they have the potential to increase learner awareness, reflection, autonomy, and goal setting (Fox 2014; Little 2005) and have both "formative and motivational value" (Williams 2014, p. 565). Some (Morris and Cooke-Plagwitz 2008; Williams 2014) argue that portfolio approaches have the potential to reduce reliance on numbers and grades by providing a more meaningful and informative, evidence-intensive alternative for profiling individual learning.

Portfolios are also increasingly evident in teacher preservice and in-service or professional development contexts as well (e.g., Delandshere and Petrosky 1998) and are playing a prominent role in course or program evaluation (e.g., Black et al. 2011).

2. *Technologically enhanced approaches to alternative assessment*

Technological advancement has permitted the use of many alternative assessment approaches which promote learning and/or provide evidence of achievement, competence, or ability. For example, electronic or e-portfolios (i.e., digital repositories of texts, presentations, videos, etc.) are increasingly used not only to support and document the learning and achievement of students but also for their teachers' preservice preparation and in-service professional development (e.g., Mansvelder-Longayroux et al. 2007). As technologically enhanced online learning activities (e.g., blogs, wikis, discussion forums) continue to develop, so too will the alternatives to assess them.

Given current concerns for educational quality and accountability, e-portfolios are increasingly being used for program evaluation (e.g., Morris and Cooke-Plagwitz 2008; Williams 2014). Mining the data housed on such e-portfolios has led to the emerging field of learning analytics in education (Williams 2014), which is "generating objective, summative reports for course certification, while at the same time providing formative assessment to personalise the student experience" (p. 5).

Hargreaves et al. (2002) argue for an interactive and collaborative assessment system in which all stakeholders – learners, teachers, parents, schools, policy

makers, etc. – make "assessment, learning, and teaching more technologically sophisticated, more critical and empowering, more collaborative and reflective, than they have ever been" (p. 92). The many new approaches to diagnostic assessment may be viewed as examples of interactive and collaborative systems of assessment that provide individualized feedback on language (and other knowledge, competencies, characteristics), through the generation of individual learning profiles. Such profiles can be the locus of student and teacher collaboration, provide an evidence-driven source of information for pedagogical interventions, and accumulate evidence over time, to monitor progress, support, and development.

E-assessment can provide an efficient and inexpensive alternative to other forms of assessment. For example, online assessment (e.g., electronic tasks and e-raters) can be used in large-scale assessment contexts, in which performance and a detailed learner profile are synchronous. It can incorporate multimodal and 3D virtual learning spaces where test-taker surrogates (avatars) interact, learn, and are assessed in digital learning spaces.

## Problems and Difficulties

While alternative assessment approaches have become more pervasive over the past decade, concerns have been raised about their validity and reliability. Although such approaches (and portfolios in particular) continue to be viewed as more *trustworthy* ( Smith and Tillema 2003), largely because of the array of evidence that can be taken into account in providing feedback, reaching a decision, or making a judgment, many researchers (e.g., Hargreaves et al. 2002; Smith and Tillema 2003) warn that it is a mistake to assume that such assessment is inherently more valid or more ethical. They point out that it is *how* alternative assessment is used that defines its character and potential; it is more than a matter of form or format (Fox 2014).

Hamp-Lyons and Condon (2009) note that "increased accuracy is not an inherent virtue of . . . [such] assessment" (p. 327). They explain that assessors must review and evaluate more evidence, varied texts, in differing numbers, across a range of genres and assignment contexts. From the perspective of measurement specialists, this makes consistent and dependable scoring much more difficult. Kane et al. (1999) have argued that if reliability is lost, the relevance of the performance is questionable, because it cannot be measured as a result. Kane et al. (1999) suggest that the goal for alternative assessment is to "achieve relevance without sacrificing too much reliability/generalizability" (p. 12).

Indeed, it is the unique and varying nature of the evidence collected as a result of alternative assessment approaches that challenges raters. When alternative assessment has been used in large-scale, high-stake contexts, such as Vermont's portfolio assessment program, there were difficulties elaborating scoring guides that had an appropriate level of specificity (i.e., were neither too brief nor too detailed) (Koretz et al. 1994).

Many of the problems and issues with alternative assessment approaches arise in the process of their implementation. A number of researchers (e.g., Fox 2014;

Hamp-Lyons and Condon 2000) who have examined the implementation of alternative assessment approaches identify ambiguity regarding purpose, use, and benefits as a primary source of difficulty. For example, Smith and Tillema (2003) consider problems arising from the implementation of portfolio assessment approaches. They point out that "[d]espite a wide array of purposes for the portfolio, including summative as well as formative assessment, selection, promotion, appraisal, reflective learning and professional development, there are many tensions and obscurities involved in portfolio use (p. 626)." They examine the range of "definitions and interpretations" (p. 625) and consider the need for increased clarity with regard to portfolio use and purpose.

Koretz et al. (1994) stress the need for "realistic expectations" (p. 11). They argue that assertions that alternative assessment is more *authentic* and *real* than other traditional forms of assessment are unrealistic. Or as Hargreaves et al. (2002) point out, "few things are more contrived and less authentic than authentic assessment, where there is a constant sorting, sifting, and reflecting on one's achievements in a portfolio, assessing one's peers using complex grids of criteria, or engaging in stage-managed three-way interviews with parents and students" (pp. 89–90).

Reflecting on the large-scale, state-wide implementation of alternative assessment (i.e., portfolio assessment) in Vermont, Koretz et al. (1994) argue it is essential to "acknowledge the large costs in time, money, and stress" (p. 11) that such alternative assessment approaches require. Increased workload and time are frequently cited problems associated with alternative assessment approaches (Hargreaves et al. 2002). Portfolio assessment, performance assessment, anecdotal comments on learning progress, responses to learning logs, communication with stakeholders, development of learning profiles, and conferences not only require a great deal of time but also levels of expertise that may not be sufficiently developed to sustain implementation (Fox 2014). Introductions to alternative assessment are often limited to a series of pre-implementation workshops. Typically, ongoing support, increased time to plan and collaborate, and supplementary resources are limited, and all of the pressures and expectations of day-to-day work are ongoing and unrelenting.

Hargreaves et al. (2002) discuss the increased workload imposed on teachers who lament their experience in *portfolio prisons*, which demand so much more of their time. The implementation of a portfolio assessment strategy may be undermined (e.g., Fox 2014) if there is insufficient support for teachers (human and material), who may resent or may not fully understand, how to use portfolio assessment in their classrooms.

Students may also resist or subvert the potential learning benefits of the portfolio, because they (and their parents) do not fully understand or appreciate the new focus on learning process and feedback (as opposed to the traditional focus on scores, grades, and ranks). Students do not generally have sufficient experience with alternative assessment to automatically value it and may instead fake or pretend to reflect, redraft, and revise (Carlson and Albright 2012; Spence-Brown 2001), rather than genuinely reflecting, revising, and learning.

Alternative assessment approaches in general and portfolio assessment in particular necessitate careful articulation of program goals and values (and the

development of a critical understanding of them across stakeholders) if such assessment is to be better than other traditional assessment approaches. Teachers need time to develop a deep understanding of how alternative assessment may potentially enhance motivation, engagement, and learning. Unless teachers understand *why* alternative assessment approaches support learning, they will be unable to develop their students' understanding and participation.

Teachers may not be able to connect ongoing assessment in support of individual learning with curricular goals or to effectively communicate information about development and growth to parents and other stakeholders, who expect to see (and who understand) grades and scores.

Quality and accountability messages from policy makers may also confuse and undermine implementation of an alternative assessment initiative if the initiative is perceived as a covert attempt to monitor, compare, and control.

In sum, although alternative assessment approaches continue to be used in many educational contexts, their implementation remains problematic and fraught with challenges (e.g., Fox 2014). The danger that alternative assessment will increase levels of surveillance and control (given the vast and permanent digital repositories that technological resources such as e-portfolios allow), is also widely discussed in the research literature (Carlson and Albright 2012).

## Future Directions

Technological advancement will continue to extend alternative assessment approaches. The e-raters, e-portfolios, and 3D virtual assessment tasks of today are only initial examples of what technology may contribute in future to e-alternatives in assessment. The roles of corpus analysis, learning analytics, and the mining of big data for assessment purposes will continue to expand.

Indeed, technology will potentially revolutionize the ways in which we engage in assessment. Even now, scientists at the Max Planck Institute, using the Oculus Rift headset and a room full of cameras, are able to simulate a virtual world in which a human participant, wearing the wireless headset, moves through physical chairs on a virtual airplane or walks through a grove of trees outside a virtual Italian villa. As such technologies develop, the potential to simulate contexts and collect evidence over time of language performance, expertise, and development will increase the reliability and validity of assessment. Elsewhere, neuroscientists and neurolinguists in collaboration with language testers are currently using functional magnetic resonance imaging (fMRI) to assess brain changes that are related to language proficiency development (e.g., Fox and Hirotani 2016). As scanning technologies are refined and improved, evidence of brain changes, which are related to increasing language proficiency, will help us to better understand the incremental cognitive changes that underlie or accompany language proficiency development. Although such approaches may be challenged as invasive, ultimately they will enhance our understanding of the cognitive changes that occur as part of second-language acquisition and not only increase the potential accuracy of construct definition in

language testing but also the validity and reliability of rating practices, outcomes, and their interpretation.

In sum, the array of alternative assessment approaches will continue to be enhanced by technology. In all cases, however, we should be guided by the sage advice of Merrill Swain, to "bias for the best" in language testing and assessment.

## Cross-References

- ▶ Dynamic Assessment
- ▶ Language Assessment Literacy
- ▶ Task and Performance-Based Assessment
- ▶ The Common European Framework of Reference (CEFR)

## Related Articles in the Encyclopedia of Language and Education

Ulrike Jessner-Schmid: Multicompetence Approaches to Language Proficiency Development in Multilingual Education. In Volume: Bilingual and Multilingual Education

Mary Toulouse, Michelle Geoffrion-Vinci: Electronic Portfolios in Foreign Lan guage Learning. In Volume: Second and Foreign Language Education

Per Urlaub: Multiliteracies and Curricular Transformation. In Volume: Second and Foreign Language Education

## References

Bailey, K. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Pacific Grove: Heinle & Heinle.

Black, P., & Wiliam, D. (2006). *Inside the black box: Raising standards through classroom assessment*. London: Granada Learning.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). 'Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles Policy & Practice, 18*(4), 451–469.

Brown, J. D., & Hudson, T. D. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*, 653–675.

Carlson, D., & Albright, J. (2012). *Composing a care for the self: A critical history of writing assessment in secondary English education*. Rotterdam: Sense Publishers.

Cheng, L., & Fox, J. (2013). Review of doctoral research in language assessment in Canada (2006–2011). *Language Teaching: Surveys and Studies, 46*, 518–544. doi:10.1017/S0261444813000244.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14–24.

Elbow, P., & Belanoff, P. (1986). Portfolios as a substitute for proficiency exams. *CCC, 37*, 336–339.

Fox, J. (2014). Portfolio based language assessment (PBLA) in Canadian immigrant language training: Have we got it wrong? *Contact Special Research Symposium Issue, 40*(2), 68–83.

Fox, J., & Hirotani, M. (2016). Detecting incremental changes in oral proficiency in neuroscience and language testing: Advantages of Interdisciplinary collaboration. In V. Arayadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 89–120). Cambridge, UK: Cambridge Scholars Press.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355–392.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Cresskill: Hampton Press.

Hamp-Lyons, L., & Condon, W. (2009). Questioning assumptions about portfolio-based assessment. *CCC 44*(2) (1993), 176–190, In B. Huot & P. O'Neill (Eds.), *Assessing writing* (pp. 315–329). Boston: Bedford/St. Martin's.

Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal, 30*(1), 69–95.

Huerta-Macías, A. (1995). Alternative assessment: Responses to commonly asked, questions. *TESOL Journal, 5*(1), 8–11.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lacelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55–75.

Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? *Language Assessment Quarterly, 4*(3), 257–278.

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335–359.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgments in the assessment process. *Language Testing, 22*(3), 321–336.

Lynch, B. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351–372.

Lynch, B., & Shaw, P. (2005). Portfolios, power, and ethics. *TESOL Quarterly, 39*(2), 263–297.

Mansvelder-Longayroux, D., Beijaard, D., & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. *Teaching and Teacher Education, 23*(1), 47–62.

Maslovaty, N., & Kuzi, E. (2002). Promoting motivational goals through alternative or traditional assessment. *Studies in Educational Evaluation, 28*, 199–222.

McNamara, T. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Language testing and assessment, Vol. 7, pp. 131–139). Boston: Kluwer Academic Publishers.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Morris, M., & Cooke-Plagwitz, J. (2008). One undergraduate Spanish program's experience with program assessment: The role of e-portfolios. *Hispania, 91*(1), 176–187.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing, 18*(4), 429–462.

Shohamy, E. (1996). Language testing: Matching assessment procedures with language knowledge. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 143–159). Boston: Kluwer.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391.

Smith, K., & Tillema, H. (2003). Clarifying different types of portfolio use. Assessment & Evaluation in Higher Education, *28*(6), 625–648.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*(4), 463–481.

Valdés, G., & Figueroa, R. A. (1994). Bilingualism and testing: A special case of bias. Ablex Publishing.

Williams, P. (2014). Squaring the circle: A new alternative to alternative assessment. *Teaching in Higher Education, 19*(5), 565–577.

# Utilizing Technology in Language Assessment

Carol A. Chapelle and Erik Voss

**Abstract**

This entry presents an overview of the past, present, and future of technology use in language assessment, also called computer-assisted language testing (CALT), with a focus on technology for delivering tests and processing test takers' linguistic responses. The past developments include technical accomplishments that contributed to the development of computer-adaptive testing for efficiency, visions of innovation in language testing, and exploration of automated scoring of test takers' writing. Major accomplishments include computer-adaptive testing as well as some more transformational influences for language testing: theoretical developments prompted by the need to reconsider the constructs assessed using technology, natural language-processing technologies used for evaluating learners' spoken and written language, and the use of methods and findings from corpus linguistics. Current research investigates the comparability between computer-assisted language tests and those delivered through other means, expands the uses and usefulness of language tests through innovation, seeks high-tech solutions to security issues, and develops more powerful software for authoring language assessments. Authoring language tests with ever changing hardware and software is a central issue in this area. Other challenges include understanding the many potential technological influences on test performance and evaluating the innovations in language assessment that are made possible through the use of technology. The potentials and challenges of technology use in language testing create the need for future language testers with a strong background in technology, language testing, and other areas of applied linguistics.

C.A. Chapelle (✉)
Applied Linguistics, Department of English, Iowa State University, Ames, IA, USA
e-mail: carolc@iastate.edu

E. Voss
NU Global, Northeastern University, Boston, MA, USA
e-mail: e.voss@neu.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_10

**Keywords**
Assessment • Computer • Technology • Testing

## Contents



## Introduction

Technology is often associated with efficiency. Accordingly, applied linguists might consider technology in language assessment in terms of how it streamlines the testing process. Indeed, much progress can be identified with respect to this worthwhile goal, as many language tests today are delivered by computer to increase efficiency. An equally important strand of language assessment concerns the relationship of language assessment to language learning, language teaching, and knowledge within the field of applied linguistics. The story of technology in language assessment needs to encompass both the efficiency of technical accomplishments and the ways that these tests intersect with other factors in the educational process for language learners. Technology can include a broad range of devices used in the testing process, from recording equipment, statistical programs, and databases to programs capable of language recognition (Burstein et al. 1996). However, here the focus will be on the use of computer technology for delivering tests and processing test takers' linguistic responses because these are the practices with the most direct impact on test takers and educational programs. The use of computer technology in language assessment is referred to as computer-assisted language assessment or computer-assisted language testing (CALT), two phrases that are used interchangeably.

## Early Developments

Early developments in computer-assisted language assessment consisted of a few demonstration projects and tests used in university language courses. Many of these were reported in two edited collections, *Technology and Language Testing* (Stansfield 1986) and *Computer-Assisted Language Learning and Testing: Research Issues and Practice* (Dunkel 1991), but others had been published as journal articles. Three important themes were prevalent in this early work.

One was the use of a psychometric approach called item response theory (Hambleton et al. 1991), which provides a means for obtaining robust statistical data on test items. These item statistics, obtained from pretesting items on a large group of examinees, are used as data by a computer program to help select appropriate test questions for examinees during test taking. Item response theory, which offers an alternative to calculation of item difficulty and discrimination through classical true score methods, entails certain assumptions about the data. The use of these methods, the assumptions they entail, and the construction and use of the first computer-adaptive tests comprised the major preoccupation of the language testers at the beginning of the 1980s. This was also the time when the first microcomputers were within reach for many applied linguists. Most of the papers in the early edited volumes in addition to journal articles (e.g., Larson and Madsen 1985) focused on issues associated with computer-adaptive testing. For example, reporting on a computer-adaptive test developed to increase efficiency of placement, Madsen (1991, p. 245) described the goal as follows: "intensive- English directors confirmed that the instrument they needed was an efficient and accurate ESL proficiency test rather than a diagnostic test." He describes the results of the research and development efforts in terms of the number of items required for placement, the mean number of items attempted by examinees, the mean amount of time it took students to complete the test, and students' affective responses to taking the test on the computer.

Other early developments appeared in a few papers exploring possibilities other than adaptivity, which were presented through the use of technology. The first issue of *Language Testing Update* at Lancaster University entitled "Innovations in language testing: Can the micro- computer help?" addressed the many capabilities of computers and how these might be put to use to improve language assessment for all test users, including learners (Alderson 1988). A paper in *CALICO Journal* at that time raised the need to reconcile the computer's capability for recording detailed diagnostic information with the test development concepts for proficiency testing, which are aimed to produce good total scores (Clark 1989). A few years later, Corbel (1993) published a research report at the National Centre for English Language Teaching and Research at Macquarie University, *Computer-Enhanced Language Assessment*, which also raised substantive questions about how technology might improve research and practice in language teaching and testing.

This early work expressed a vision of the potential significance of technology for changes and innovation in second language assessment, an agenda-setting collection of questions. However, the technology agenda for language assessment requires considerable infrastructure in addition to cross-disciplinary knowledge dedicated to problems in language assessment. At this time, decision-makers at the large testing companies, where such resources resided, apparently did not see technology-based assessment as a practical reality for operational testing programs. Instead, discussion of just a few innovative projects produced in higher education appeared (Marty 1981).

Significant advances involving computer recognition of examinees' constructed responses remained in research laboratories and out of reach for assessment practice

(Wresch 1993). This frustrating reality coupled with technical hardware and software challenges and the intellectual distance between most applied linguists and technology resulted in a slow start. By 1995, many applied linguists were voicing doubts and concerns about the idea of delivering high-stakes language tests by computer, fearing that the negative consequences would far outweigh any advantages. As it turned out, however, the technologies affecting language assessment did not wait for the approval and support of applied linguists. By the middle of the 1990s, many testing programs were beginning to develop and use computer-assisted language tests.

## Major Contributions

The rocky beginning for technology in language assessment is probably forgotten history for most test users, as major contributions have now changed the assessment landscape considerably. Language test developers today at least consider the use of technology as they design new tests. Test takers and score users find online tests to be the norm like other aspects of language learning curricula and tools used in other facets of life. Contributions are complex and varied, but they might be summarized in terms of the way that technology has advanced language testing in four ways.

First, computer-adaptive testing has increased the efficiency of proficiency and placement testing. Many computer-adaptive testing projects have been reported regularly in edited books (i.e., Chalhoub-Deville 1999, and the ones cited earlier) and journal articles (e.g., Burston and Monville-Burston 1995). By evaluating examinees' responses immediately as they are entered, a computer-adaptive test avoids items that are either too easy or too difficult; such items waste time because they provide little information about the examinee's ability. In addition to creating efficient tests, these projects have raised important issues about the way language is measured, the need for independent items, and their selection through an adaptive algorithm. One line of research, for example, examines the effects of various schemes for adaptivity on learners' affect and test performance (Vispoel et al. 2000). Another seeks strategies for grouping items in a manner that preserves their context to allow several items to be selected together because they are associated with a single reading or listening passage. Eckes (2014), for example, investigated testlet effects in listening passages for a test of German as a foreign language.

Second, technology has prompted test developers to reconsider the constructs that they test. One example is the use of multimedia in testing listening comprehension. In the past, the testing of listening comprehension was limited to the examiner's oral presentation of linguistic input, either live or prerecorded, to a room full of examinees. Such test methods can be criticized for their failure to simulate listening as it occurs in many contexts, where visual cues are also relevant to interpretation of meaning. The use of multimedia provides test developers with the opportunity to contextualize aural language with images and to allow examinees to control their test-taking speed and requests for repetition. This option for construction of a test, however, brings interesting research questions about the nature of listening and the

generalizability of listening across different listening tasks. Some of these questions are being explored in research on integrated tasks, which combine requirements for reading, writing, and speaking, for example. In this research, eye-tracking technology has proven useful for investigating how test takers interact with such tasks (Suvorov 2015).

Another example is the assessment of low-stakes dialogic speaking using Web cameras and videoconferencing software. Video simulates an interview in person with affordances for nonverbal skills, which are not available in monologic speech samples. For example, Kim and Craig (2012) found that linguistic performance on face-to-face English proficiency interviews was similar to performance on interviews conducted using videoconferencing software. The nonlinguistic cues such as gestures and facial expressions, however, were absent or difficult to see because of the small screen size. Advances in computer technology in research settings have made possible automatic assessment of dialogic oral interactions that include nonverbal communication. A computerized conversation coach developed at Massachusetts Institute of Technology, for example, provides summaries of oral and facial expressions such as head nodding and smiling through automated analysis in addition to speech recognition and prosody analysis in a simulated conversation (Hoque 2013). A third example is the use of actuators and sensors that sense changes in human emotion and mood, for instance, when a test taker is nervous during an oral interview. Although these technologies are not yet integrated in testing, Santos et al. (2016) are exploring the use of ambient intelligence to provide real-time natural interaction through visual, audio, and tactile feedback by a computer in response to changes in a learner affective state during a mock interview. These technological capabilities integrated into future assessments will allow test developers to assess both verbal and nonverbal aspects of speaking and in doing so will constantly require rethinking and investigating the construct meaning.

Third, natural language-processing technologies are being used for evaluating learners' spoken and written language. One of the most serious limitations with large-scale testing in the past was the over-reliance on selected-response items, such as multiple choice. Such items are used because they can be machine scored despite the fact that language assessment is typically better achieved if examinees produce language as they need to do in most language-use situations. Research on natural language processing for language assessment has recently yielded technologies that can score learners' constructed linguistic responses as well. A special issue of *Language Testing* in 2010 describes the research in this area and points to the use of these technologies in operational testing programs, typically for producing scores based on an evaluation of a response. Such evaluation systems are also being put to use for low-stakes evaluation and feedback for students' writing (Chapelle et al. 2015). Such work has advanced farther for responses that are written than those that are spoken.

Fourth, corpus linguistics is used to inform the design and validation of language assessments (Park 2014). A corpus can consist of texts produced by language learners or a collection of texts representing the target language-use domain relevant to score interpretation. Learner corpora are used by test developers to identify

criterial linguistic features that appear in learners' language at particular stages of development. Such features can be used to produce descriptors for evaluating learners' constructed responses or to investigate the language elicited from particular test tasks.

Corpora representing the target language-use domain can be used to identify lexical, structural, and functional content that characterizes a particular language domain. One purpose of defining the domain is to ensure that test tasks are modeled on tasks that test takers will perform in the target domain (e.g., Biber 2006). Such an investigation can result in selection of specific linguistic features for test items as Voss (2012) did by sampling collocations from a corpus of academic language. In this case, the corpus was also used to verify frequent and possible collocations to inform a partial-credit scoring procedure. Similarly, reading and listening passages can be selected or developed with appropriate difficulty levels based on the frequency of lexis in the passage aligned with characteristics identified in corresponding proficiency levels. Using frequency and sentence length data, for example, standardized Lexile® scores for reading passages are used to complement assessment results with level-appropriate instruction and reading ability levels (Metametrics 2009). The systematicity and empirical basis of linguistic analysis during test development are an important part of the evidence in a validity argument for the test score interpretations.

These technical advances in test methods need to be seen within the social and political contexts that make technology accessible and viable to test developers, test takers, and test users. Not long ago most test developers felt that the operational constraints of delivering language tests by computer may be insurmountable. Today, however, many large testing organizations are taking advantage of technical capabilities that researchers have been investigating for at least the last 20 years. As computer-assisted language assessment has become a reality, test takers have needed to reorient their test preparation practices to help them prepare.

## Work in Progress

The primary impetus for using technology in language assessment was for many years to improve the efficiency of testing practices and thus much of the work in progress has centered on this objective. Research is therefore conducted when testing practices are targeted for replacement by computer-assisted testing for any number of reasons such as an external mandate. The objective for research in these cases is to demonstrate the equivalence of the computer-assisted tests to the existing paper-and-pencil tests. For example, such a study of the Test of English Proficiency developed by Seoul National University examined the comparability of computer-based and paper-based language tests (CBLT and PBLT, respectively). Choi et al. (2003) explained the need for assessing comparability in practical terms: "Since the CBLT/CALT version of the [Test of English Proficiency] TEPS will be used with its PBLT version for the time being, comparability between PBLT and CBLT is crucial

if item statistics and normative tables constructed from PBLT are to be directly transported for use in CBLT" (Choi et al. 2003, p. 296). The study, which used multiple forms of analysis to assess comparability of the constructs measured by the two tests, found support for similarity of constructs across the two sets of tests, with the listening and grammar sections showing the strongest similarities and the reading sections showing the weakest.

In addition to the practical motivation for assessing similarity to determine whether test scores can be interpreted as equivalent, there is an important scientific question to be investigated as well: what important construct-relevant differences in language performance are sampled when technology is used for test delivery and response evaluation. Unfortunately, few studies have tackled this question (Sawaki 2001). The use of technology for test delivery is frequently a decision that is made before research, and therefore the issue for practice is how to prepare the examinees sufficiently so that they will not be at a disadvantage due to lack of computer experience. For example, Taylor et al. (1999) gave the examinees a tutorial to prepare them for the computer-delivered items before they investigated the comparability of the computer-based and the paper-and-pencil versions of test items for the (TOEFL). In this case, the research objective is to demonstrate how any potential experience-related difference among test takers can be minimized. The need for tutorials is disappearing as younger learners grow up with computer technology. Computer and language literacy develop together as the use of touch-screen tablets in homes and early education is increasing (Neumann 2016). In response such new practices for literacy development, The Cambridge English Language Assessment allows young test takers to choose their preferred mode of test delivery by taking the test on a computer or on paper (Papp and Walczak 2016). The results of research investigating performance on both show that the two delivery modes were comparable, that "children are very capable of using computers, and that they especially like using iPads/tablets" (p. 168).

As technology has become commonplace in language education, researchers and developers hope to expand the uses and usefulness of language tests through innovation. For example, the DIALANG project, an an Internet-based test, developed shortly after the advent of the Web (Alderson 2005), was intended to offer diagnostic information to learners to increase their understanding of their language learning. Whereas DIALANG was intended to have extensive impact on language learners due to its accessibility on the Web, other assessments aimed at learning appear in computer-assisted language learning materials. Longman English Interactive (Rost 2003), for example, includes assessments regularly throughout the process of instruction to inform learners about how well they have learned what was taught in each unit. Such assessments, which also appear in many teacher-made materials, use technology to change the dynamic between test takers and tests by providing learners a means for finding out how they are doing, what they need to review, and whether they are justified in their level of confidence about their knowledge. These same ideas about making assessment available to learners through the delivery of low-stakes assessment are migrating to the next generation of technologies.

For example, Palomo-Duarte et al. (2014) describe a low-stakes test of vocabulary that learners can take on their smartphones by downloading an app. Also, designed to meet student demand, many apps have been created to accompany language learning or as practice test for standardized language tests such as TOEFL and IELTS.

For high-stakes testing, in contrast, lack of adequate security poses a thorny problem for assessment on mobile platforms. However, because mobile devices with multimedia capabilities and Internet access are becoming so commonplace, the development of low-cost, large-scale, high-stakes language tests with multimodal interaction is enticing. For example, two universities in Spain are exploring the delivery of the Spanish University Entrance Examination on mobile devices (García Laborda et al. 2014). The mobile-enhanced delivery of the Spanish test includes assessment of grammar, reading, writing, listening, and speaking with a combination of (automated rating and responses assessed later by human raters). Currently, such devices are best suited for listening and speaking tasks because small screen sizes on mobile phones make appropriate reading tasks difficult to construct. Technological limitations also affect the expected written responses that can be requested of test takers. Producing written language on a smartphone entails a number of fundamental differences from writing at a keyboard, and therefore, the device needs to be considered carefully in the design of test tasks. Smartphone testing issues are undoubtedly entering into mainstream language testing because their reach extends even beyond that of the Internet. In physical locations where the Internet connection is slow or nonexistent, language tests have been administered using the voice and SMS texting technologies of mobile phones (Valk et al. 2010). Delivery of assessments to students in remote areas is possible with these platforms even if supplemental paper-based materials are necessary.

All of this language testing development relies on significant software infrastructure, and therefore another area of current work is the development of authoring systems. Due to limitations in the existing authoring tools for instruction and assessment, most language-testing researchers would like to have authoring tools intended to address their testing goals directly, including the integration of testing with instruction, analysis of learners' constructed responses, and capture and analysis of oral language. As such, capabilities are contemplated for authoring tools, as are new ways for conceptualizing the assessment process. Widely used psychometric theory and tools were developed around the use of dichotomously scored items that are intended to add up to measure a unitary construct. The conception of Almond et al. (2002) underlying their test authoring tools reframes measurement as a process of gathering evidence (consisting of test takers' performance) to make inferences about their knowledge and capabilities. The nature of the evidence can be, but does not have to be, dichotomously scored items; it can also be the results from a computational analysis of learners' production. Inferences can be made about multiple forms of knowledge or performance. The emphasis on evidence and inference underlies plans for developing authoring tools for computer-assisted testing that can include a variety of types of items and can perform analysis on the results that are obtained – all within one system.

## Problems and Difficulties

With the intriguing potentials apparent in current work, many challenges remain, particularly in view of the changing technologies. Testing programs need to have built-in mechanisms for updating software, hardware, and technical knowledge of employees. Large testing companies with the most resources may be the most able to keep up with changes. To some extent they have done so by increasing fees for those using their tests. In some cases costs are borne by language programs, but in many other cases, the costs are passed on to those who are least able to pay – the test takers themselves. Small testing organizations, publishing companies for whom testing is just one part of their overall profile, as well as school-based testing programs have to rely on strategic partnerships to combine expertise, limited resources, and technologies. Navigation of these waters in a quickly changing environment requires exceptionally knowledgeable leadership.

Challenges that may be less evident to test users are those that language-testing researchers grapple with as they attempt to develop appropriate tests and justify their use for particular purposes. As Bachman (2000, p. 9) put it, "the new task formats and modes of presentation that multimedia computer-based test administration makes possible raise all of the familiar validity issues, and may require us to redefine the very constructs we believe we are assessing." For example, Chapelle (2003) noted that in a computer-assisted reading test, the test tasks might allow the test takers access to a dictionary and other reading aids such as images. In this case, the construct tested would be the ability to read with strategic use of online help. The reading strategies entailed in such tasks are different from those used to read when no help is available, and therefore the definition of strategic competence becomes critical for the construct assessed. Should test takers be given access to help while reading on a reading test? One approach to the dilemma is for the test developer to decide whether or not access to help constitutes an authentic task for the reader. In other words, if examinees will be reading online with access to help, such options should be provided in the test as well. However, the range of reading tasks the examinees are likely to engage in is sufficiently large and diverse to make the authentic task approach unsatisfactory for most test uses. The reading construct needs to be defined as inclusive of particular strategic competencies that are required for successful reading across a variety of contexts.

A second example of how technology intersects with construct definition comes from tests that use natural language processing to conduct detailed analyses of learners' language. Such analyses might be used to calculate a precise score about learners' knowledge or to tabulate information about categories of linguistic knowledge for diagnosis. In either case, if an analysis program is to make use of such information, the constructs assessed need to be defined in detail. A general construct definition such as "speaking ability" does not give any guidance concerning which errors and types of disfluencies should be considered more serious than others, or which ones should be tabulated and placed in a diagnostic profile. Current trends in scoring holistically for overall communicative effectiveness circumvent the need for taking a close linguistic look at constructed responses. One of the few studies to

grapple with this issue (Coniam 1996) pointed out the precision afforded by the computational analysis of the learners' responses far exceeded that of the construct of listening that the dictation test was measuring. To this point assessment research has not benefited from the interest that second language acquisition researchers have in assessing detailed linguistic knowledge; it remains a challenge (Alderson 2005).

Another challenge that faces language-testing researchers is the need to evaluate computer-assisted language tests. As described earlier, current practices have focused on efficiency and comparability. However, one might argue that the complexity inherent in new forms of computer-assisted language assessment should prompt the use of more sensitive methods for investigating validity. When the goal of test development is to construct a more efficient test, then efficiency should clearly be part of the evaluation, but what about computer-assisted tests that are intended to provide more precise measurement, better feedback to learners, or greater accessibility to learners? If the scores obtained through the use of natural language-processing analysis are evaluated by correlating them with scores obtained by human raters or scores obtained with dichotomously scored items (e.g., Henning et al. 1993), how is the potential additional value of the computer to be detected?

In arguing for evaluation methods geared toward computer-assisted language tests, some language-testing researchers have focused on interface issues (Fulcher 2003) – an important distinction for computer-assisted tests. It seems that the challenge is to place these interface issues within a broader perspective on validation that is not overly preoccupied by efficiency and comparability with paper-and-pencil tests. Chapelle et al. (2003), for example, frame their evaluation of a Web-based test in broader terms, looking at a range of test qualities. Chapelle and Douglas (2006) suggest the continued need to integrate the specific technology concerns into an overall agenda for conceptualizing validation in language assessment that includes the consequences of test use. Technology reemphasizes the need for researchers to investigate the consequences of testing. Such consequences might include benefits such as raising awareness of the options for learning through technology.

## Future Directions

These two sets of challenges – the obvious ones pertaining to infrastructure and the more subtle conceptual issues evident to language-testing researchers – combine to create a third issue for the field of applied linguistics. How can improved knowledge about the use of technology be produced and disseminated within the profession? What is the knowledge and experience that graduate students in applied linguistics should attain if they are to contribute to the next generations of computer-assisted language tests? At present, it is possible to identify some of the issues raised through the use of technology that might be covered in graduate education, but if graduate students are to dig into the language-testing issues, they need to be able to create and experiment with computer-based tests.

Such experimentation requires authoring tools that are sufficiently easy to learn and transportable beyond graduate school. Commercial authoring tools that are

widely accessible are not particularly suited to the unique demands of language assessment such as the need for linked items, the evaluation of learners' oral and written production, and the collection of spoken responses. As a consequence, many students studying language assessment have no experience in considering the unique issues that these computer capabilities present to language testing. In a sense, the software tools available constrain thinking about language assessment making progress evolutionary rather than revolutionary (Chapelle and Douglas 2006).

More revolutionary changes will probably require graduate students educated in language testing in addition to other areas of applied linguistics. For example, students need to be educated in corpus linguistics to conduct appropriate domain analyses as a basis for test development (e.g., Biber 2006). Education in second-language acquisition is needed too for students to use learner corpora for defining levels of linguistic competence (Saville and Hakey 2010). Education in world Englishes is needed to approach issues of language standards (Mauranen 2010). These and other aspects of applied linguistics appear to be critical for helping to increase the usefulness of assessment throughout the educational process, strengthen applied linguists' understanding of language proficiency, and expand their agendas for test validation.

## Cross-References


- ▶ Cognitive Aspects of Language Assessment
- ▶ Using Portfolios for Assessment/Alternative Assessment
- ▶ Utilizing Accommodations in Assessment


## Related Articles in the Encyclopedia of Language and Education

Rémi A. van Compernolle: Sociocultural Approaches to Technology Use in Language Education. In Volume: Language, Education and Technology

Kevin M. Leander, Cynthis Lewis: Literacy and Internet Technologies. In Volume: Literacies and Language Education

John Thorne: Technologies, Language and Education: Introduction. In Volume: Language, Education and Technology

Paula Winke, Daniel R. Isbell: Computer-Assisted Language Assessment. In Volume: Language, Education and Technology

## References


Alderson, J. C. (1988). *Innovations in language testing: Can the microcomputer help? Special Report No 1 Language Testing Update*. Lancaster: University of Lancaster.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning and Assessment, 1*(5). Available from http://www.jtla.org

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Burstein, J., Frase, L., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics, 16*, 240–260.

Burston, J., & Monville-Burston, M. (1995). Practical design and implementation considerations of a computer-adaptive foreign language test: The Monash/Melbourne French CAT. *CALICO Journal, 13*(1), 26–46.

Chalhoub-Deville, M. (Ed.). (1999). *Development and research in computer adaptive language testing*. Cambridge: University of Cambridge Examinations Syndicate/Cambridge University Press.

Chapelle, C. A. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology.* Amsterdam: John Benjamins Publishing.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409–439.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Diagnostic assessment with automated writing evaluation: A look at validity arguments for new classroom assessments. *Language Testing, 32*(3), 385–405.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295–320.

Clark, J. L. D. (1989). Multipurpose language tests: Is a conceptual and operational synthesis possible? In J. E. Alatis (Ed.), *Georgetown university round table on language and linguistics. Language teaching, testing, and technology: Lessons from the past with a view toward the future* (pp. 206–215). Washington, DC: Georgetown University Press.

Coniam, D. (1996). Computerized dictation for assessing listening proficiency. *CALICO Journal, 13*(2–3), 73–85.

Corbel, C. (1993). Computer-enhanced language assessment. In G. Brindley (Ed.), *Research report series 2, National Centre for English Language Teaching and Research*. Sydney: Marquarie University.

Dunkel, P. (Ed.). (1991). *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39–61.

Fulcher, G. (2003). Interface design in computer based language testing. *Language Testing, 20*(4), 384–408.

García Laborda, J. G., Magal-Royo, T. M., Litzler, M. F., & Giménez López, J. L. G. (2014). Mobile phones for Spain's university entrance examination language test. *Educational Technology & Society, 17*(2), 17–30.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage.

Henning, G., Anbar, M., Helm, C., & D'Arcy, S. (1993). Computer-assisted testing of reading comprehension: Comparisons among multiple-choice and open-ended scoring methods. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research*. Alexandria: TESOL.

Hoque, M. E. (2013). *Computers to help with conversations: Affective framework to enhance human nonverbal skills* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 0830325).

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning, 52*(3), 257–275.

Larson, J. W., & Madsen, H. S. (1985). Computer-adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal, 2*(3), 32–36.

Madsen, H. S. (1991). Computer-adaptive test of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 237–257). New York: Newbury House.

Marty, F. (1981). Reflections on the use of computers in second language acquisition. *Studies in Language Learning, 3*(1), 25–53.

Mauranen, A. (2010). Features of English as a lingua franca in academia. *Helsinki English Studies, 6*, 6–28.

Metametrics. (2009). The Lexile framework for reading. Retrieved from http://www.lexile.com

Neumann, M. M. (2016). Young children's use of touch screen tablets for writing and reading at home: Relationships with emergent literacy. *Computers & Education, 97*, 61–68.

Palomo-Duarte, M., Berns, A., Dodero, J. M., & Cejas, A. (2014). Foreign language learning using a gamificated APP to support peer-assessment. In *Proceedings of TEEM' 14: Second international conference on technological ecosystem for enhancing multiculturality*, Salamanca, Volume 1.

Papp, S., & Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (Educational linguistics, Vol. 25, pp. 139–190). New York: Springer.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*(1), 27–44.

Rost, M. (2003). *Longman English interactive*. New York: Pearson Education.

Santos, O. C., Saneiro, M., Boticario, J. G., & Rodriquez-Sanchez, M. C. (2016). Toward interactive context-aware affective educational recommendations in computer- assisted language learning. *New Review of Hypermedia and Multimedia, 22*(1), 27–57.

Saville, N., & Hakey, R. (2010). The English language profile – The first three years. *English Language Journal, 1*(1), 1–14.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*(2), 38–59.

Stansfield, C. (Ed.). (1986). *Technology and language testing*. Washington, DC: TESOL Publications.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463–483.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–274.

Valk, J. H., Rashid, A. T., & Elder, L. (2010). Using mobile phones to improve educational outcomes: An analysis of evidence from Asia. *The International Review of Research in Open and Distance Learning, 11*(1), 117–140.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement, 37*(1), 21–38.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3539432).

Wresch, W. (1993). The imminence of grading essays by computer – 25 years later. *Computers and Composition, 10*(2), 45–58.

# Cognitive Aspects of Language Assessment

Eunice Eunhee Jang

**Abstract**

The field of language assessment has expanded its focus from making inferences based on performance outcome to examining cognitive processes involving structuring information for conceptual and procedural understandings necessary for successful assessment task completion. The quality of these inferences is subject to the extent to which tasks used to elicit mental processes are successfully performed. It is equally critical that the observed mental processes do not differ significantly from those taking place in nonobservational situations. The chapter reviews research on learner cognition through an examination of reading comprehension processes in testing situations. Subsequently, it discusses major contributions and theoretical, methodological, and contextual challenges in assessing learner cognition. The chapter highlights that learners' cognitive capacity for executing tasks is best understood when they are viewed as a dynamic system. When assessment ignores such dynamic interplay of learners' multiple traits, our evidentiary reasoning about their cognitive capabilities remains incomplete. I conclude the chapter with suggestions that will extend current research with technological advances.

**Keywords**

Cognition • Metacognition • Reading comprehension • Language assessment

E.E. Jang (✉)
Department of Applied Psychology and Human Development, Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada
e-mail: eun.jang@utoronto.ca

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_11

## Contents



## Introduction

Learners' mental processes including cognitive and metacognitive processes in learning and testing situations have been a focal research topic. This reflects an appreciation of the complexities of language itself, as well as learning and assessment. For example, neuroscience research on reading shows that seemingly simple tasks, such as sounding out a word, require the orchestration of distinct areas (e.g., sensory visual processing of letters and word forms, perceptual processing of speech sounds, speech motor processing, spatial orientation (Hruby and Goswami 2011). Reading involves a construction of coherent mental representation of the text and identification of semantic relations. Reading process becomes increasingly automatized and selective, through which readers cope with limited working memories and cognitive resources (Anderson 2000; Stanovich 2001). As learners engage in a discourse-level text, they expand their word-level reading skills toward inferential and discourse-level comprehension skills. Readers activate their prior knowledge by retrieving relevant information from long-term memory, organizing information, and allocating cognitive resources (Ericsson 2003). Assessing reading ability needs to focus on the extent to which learners have achieved automaticity in processing basic encoding skills, paying selective attention to important information, and further retrieving and organizing knowledge in order to construct a coherent mental representation of the text.

Metacognitive control plays a critical role in processing textual information efficiently as it allows readers to process cognitive resources related to attention and limited working memory effectively (Carretti et al. 2009; Kendeou et al. 2014). Therefore, comprehension monitoring skills are considered the driving forces behind the development of later reading abilities (Koda 2005). Metacognition involves the ability to self-regulate one's own learning by setting goals, monitoring comprehension, executing repair strategies, and evaluating comprehension strategies. Research consistently supports the significant and positive role that metacognitive abilities play in developing language proficiency (Carrell 1998; Ehrlich et al. 1999). When

readers with poor metacognitive control are faced with reading tasks beyond their current level, they become cognitively overloaded (Sweller 1988). In this case, extraneous cognitive load is increased due to ineffective assessment tasks that fail to tap into the readers' learning capacity, resulting in negatively affecting germane cognitive load required for schema construction and automation of comprehension process (Sweller 1988). Furthermore, readers may suffer from low self-efficacy and negative emotional arousals (Pekrun et al. 2002).

Though the early development of research on learner cognition in language learning, particularly reading ability, has been integral to advancing our knowledge base, its integration into assessment has been relatively slow. In recent years, there is a significant shift toward the assessment of cognitive processes underlying language proficiency, that is, how students think and process linguistic knowledge to fulfill communicative goals and how they execute a communicative task with emerging knowledge. An assessment triangle framework by Pellegrino et al. (2001) shows the relationship among three core elements of assessment: cognition, observation, and interpretation, common across contexts and purposes (Fig. 1).

The assessment triangle highlights the importance of modeling how students develop competence in a particular subject domain, designing tasks that elicit their deep thinking and emergent understanding, and advancing measurement models for drawing rich inferences based on the cognitive evidence obtained. Evidentiary arguments for assessment require a process of reasoning from all elements of the triangle (Mislevy et al. 2003). From the vantage point of this assessment triangle, I first discuss *cognitive* theories of reading comprehension and its relationship with metacognition. Subsequently I discuss evidentiary reasoning for making diagnostic *interpretations* about cognitive processes and challenges in making *observations* about cognitive processes. The chapter concludes with areas for future research on learner cognition in language assessment.

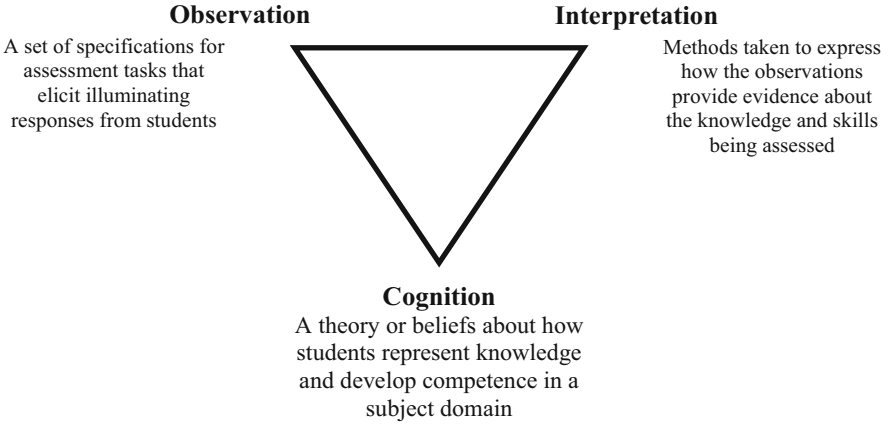


**Fig. 1** Pellegrino et al.'s assessment triangle

## Early Developments

Various cognitive theories conceptualized the process of text reading, which provided cognitive models necessary for reading assessment (Grabe 2009; Koda 2005; Kintsch and van Dijk 1978; Stanovich 1980). Early research in L1 reading process focused on bottom-up processing involving decoding small linguistic units such as phonological, lexical, and syntactic features in order to construct textual meaning (LaBerge and Samuels 1974). While information processing models contributed to our understanding of the early language development and linguistic and cognitive factors associated with reading difficulties, they were limited in grasping higher-level processing involving background knowledge and inferencing (Anderson and Pearson 1984; Carrell 1988; Goodman 1967). Goodman (1967) postulated the reading process as a reader's active participation in a set of comprehension activities involving sampling a text for graphic clues, predicting the content of text, hypothesizing text meaning, and testing and confirming the hypothesized meaning based on prior knowledge. Top-down processing reverses the order of bottom-up processing through enriching propositional meaning by making connections between the text and contextual knowledge. What differentiates these two processing models is that the bottom-up processing assumes that low-level processes are a prerequisite for high-level processes and therefore have a hierarchical relationship. On the other hand, top-down processing posits that higher-level processes can compensate for deficiencies in lower-level processes.

Interactive compensatory models (Stanovich 1980) dismiss the bottom-up model and expand the top-down model by positing that compensation for lack of knowledge can occur at any level. Stanovich (1980) argued that the effects of contextual variables on interaction with skills depend on the level of reading process with which the contextual variables are associated. He integrated the idea of interactive processing into the compensatory processing model, where making use of other knowledge sources could compensate for deficiencies at any level. For example, less skilled readers may use top-down processing by using contextual clues to compensate for poor lexical processing. However, skilled readers have no need to resort to top-down processing to handle lexical decoding. In fact, proficient readers demonstrate higher-level processes more frequently because their strength in low-level processes makes their cognitive capacity available for higher-level processes. In sum, what emerged as a consensus among the interactive theorists is that reading is an act of interplay between reader variables and text variables, which had a significant implication for assessment design and validation in later years.

Another critical early development in cognitive reading processes is the role of prior background knowledge, that is, schemata, specifically formal and content schemata in reading (Carrell 1988). Anderson and Pearson (1984) provided further evidence of how schemata function in the process of reading comprehension. Of particular importance was direct description of the role of schemata in readers' attempts to make inferences about textual information. As reading proceeds, readers

activate different schemata to connect text ideas to information retrieved from memory. Readers are able to comprehend text and infer implied meanings by making use of relevant schemata already stored in the brain. The implications of schema theory for reading are that readers are viewed to bring a variety of background knowledge and experiences to reading; consequently different readers will construct different meanings from a single text. This prompted a closer examination of retrieving information from long-term memory, organizing ideas, and allocating cognitive resources among readers with different proficiency levels (Anderson and Pearson 1984; Carrell 1988). Research suggests that poor readers tend to overly rely on either text-based processing by paying attention to decoding individual words, their lexical meanings, and syntactic structures at the sentence level to compensate for insufficient knowledge structures at the higher level of processing. On the other hand, readers with relevant content knowledge are thought to recall a text better and perform better on a discipline-specific reading comprehension test. Such early developments in research on reading processes provided insights into cognitive models of reading for making diagnostic inferences in assessment. Cognitive processing was scarce.

## Major Contributions

### Observation of Cognitive Processes

In order for assessments to capture learners' cognitive processes and knowledge structures, it requires empirical evidence for confirming the theoretical assumption regarding the extent to which the target skills intended by a test designer are congruent with actual skills learners use during assessment. Weir (2005) points out that cognitive processes cannot be fully understood from performance outcomes. Cognitive processes in assessment situations are subject to varying cognitive demands by different task types and reading purposes. Such interactive processes require more direct observations of actual processes happening, while the reader engages with the text.

Unfortunately, not all existing assessments make such assumptions explicit. The taxonomies and content specifications used to guide such approaches lack explicit cognitive theories about cognitive mechanisms underlying target domains. The substantive base that explains cognitive mechanisms is necessary before designing the test rather than post hoc, which is currently a common practice. Nichols (1994) sums up this critical point by stating "Any number of explanations are plausible when theorizing is post hoc, but fewer theories are successful in predicting results" (p. 596).

Renewed interests in reading processes and debates over the effects of testing methods on reading processes directed special attention to the potential of think-

aloud verbal protocols to elicit various types of skills and strategies. By asking readers to think aloud as they work through a series of reading comprehension questions, researchers can directly observe cognitive processes and strategies, while readers carry out tasks. Ericsson and Simon (1993) stated that when appropriately controlled, subjects' verbalization of thought processes as they engage in a problem solving activity can provide valid information about cognitive processes without distorting them.

Think-aloud verbal protocols have been used to examine the extent of congruence between skills intended by test developers and those used by students in testing and non-testing situations. Disagreements between test takers' strategies and test developers' expected skills are partly because test developers' test specifications do not provide detailed information about item-by-skill relationships, and furthermore, readers tend to use more than one skill, some of which the test developers did not predict. These observations support that test developers' intended skills do not provide a full picture of reading processes and strategies that students actually utilize during the testing situation. Evidentiary cognitive validity arguments require reasoning from direct observations of students' cognitive processes and strategies.

Learners' cognitive processes are influenced by (con)textual factors. For example, reading texts with different rhetorical organizational structures invoke different types of processing. Jang (2009b) reported that the same item types elicit different reading skills depending on the textual structures, indicating a complex interaction between textual factors and processing skills. Alderson et al. (2000) contended that competency in reading entails the ability to recognize how ideas are presented in the text and to understand authorial intentions underlying the sequence of the ideas. These findings have important implications for designing assessment. How different textual variables elicit different cognitive skills need to be reflected in the specification of skills.

In addition to textual factors, the quality of tasks used to elicit cognitive skills is a critical factor. Research shows that higher-order thinking skills, such as inferencing and summarizing skills, may be affected and altered by item position and task type. Jang (2009b) and Cohen and Upton (2007) examined students' think-aloud protocols while engaging in the same reading comprehension tasks as Jang's study and offered corroborating results. In both studies, a new item type called "prose summary" asked students to select multiple statements that best summarize the text. Jang reported that both low- and high-proficient students did not go back to the text for different reasons. The high-proficient students cumulated sufficient textual knowledge after answering all the preceding questions, whereas low-proficient students reported a lack of time to resort to the text. Similarly, Cohen and Upton (2007) reported that students used test-taking strategies because they did not need to find nor generate main statements in the text.

These research results strongly suggest that different items and text types invoke different cognitive processing, which is further differentiated by students at different proficiency levels. More research is necessary to fully understand such three-way interactions based on a large number of items eliciting the full range of cognitive skills from learners with various language proficiency levels.

## Diagnostic Interpretations of Cognitive Skills

Another major development can be attributable to cognitive diagnostic assessment. Advances in theories of cognition prompted new initiatives for diagnostic language assessment in order to provide formative diagnostic information about learners' competencies in fine-grained skills (Jang 2009a; Nichols 1994). Spolsky (1990) argued that it is testers' moral responsibility to ensure the interpretability and accuracy of test information. He suggested "profiles" that show multiple skills tested in more than one way as a more valuable reporting method. Shohamy (1992) proposed an assessment model which utilizes test results through a "detailed, innovative, and diagnostic" feedback system (p. 515).

Various cognitive diagnostic modeling approaches were developed to classify learners with distinct patterns of skill mastery and make inferences about qualities of learners' cognitive processes and knowledge structures. For example, the rule-space model uses a pattern recognition approach based on the distance between observed examinee item response patterns and a set of possible ideal response patterns. In the field of language testing, this model was applied to a short-answer listening comprehension test administered to 412 Japanese college students (Buck and Tatsuoka 1998) and to TOEFL reading subtests (Kasai 1997). The reduced Reparameterized Unified Model (r-RUM) is another cognitive diagnosis model applied to language assessment data. Jang (2005) applied the r-RUM to 2700 test takers' responses to TOEFL LanguEdge reading comprehension tests. Kim (2011) applied it to the TOEFL iBT writing test. Given few empirical studies that explicitly examined cognitive models for diagnostic interpretations, Table 1 presents cognitive attributes

**Table 1** Cognitive attributes used in cognitive diagnostic modeling

| Buck & Tatsuoka's listening attributes (1998) | Jang's reading attributes (2009a, b) | Kim's writing attributes (2011) |
|---|---|---|
| Identifying the task by determining what types of information to search for in order to complete the task | Processing context-dependent vocabulary | Fulfilling content by successfully addressing a given topic |
| Scanning fast spoken text, automatically and in real time | Processing context-independent vocabulary | Organizing and developing ideas effectively |
| Processing a large information load | Comprehending syntactic elements | Demonstrating ability to apply grammatical rules |
| Processing a medium information load | Comprehending explicit textual information | Demonstrating vocabulary knowledge with a broad range of sophisticated words |
| Processing dense information | Comprehending implicit textual information | Demonstrating correct use of English conventions |
| Using previous items to help information location | Making inferences | |
| Identifying relevant information without any explicit marker | Interpreting negatively stated information | |
| Understanding and utilizing heavy stress | Analyzing and evaluating relative importance of textual information | |
| Processing very fast text automatically | Mapping contrasting ideas | |

involving listening (Buck and Tatsuoka 1998) and writing processes (Kim 2011) in addition to reading processes (Jang 2009a, b).

The quality of diagnostic inferences from these modeling applications depends on the extent to which cognitive processing skills used are theoretically compelling and empirically justifiable. In addition, evaluating learners' competencies in microlevel skills requires a much finer-grained representation of the construct of interest. The CDA approach attempts to achieve this by evaluating individual test takers' competencies in a set of user-specified skills. The approach thus needs to be based on a substantive theory of the construct that describes processes underlying task performance. It also requires clear specifications that delineate the tasks in terms of how they elicit cognitive processes.

However, uncertainty persists in evaluating the extent to which the specified skills represent the target construct. Furthermore, all the studies reviewed above retrofitted the CDA approaches to existing proficiency tests, which might not have been designed with explicit specifications of fine-grained skills. Insufficient observations due to limited items per skill would increase inaccuracy in diagnostic classifications. It is also possible that important skills might be excluded from modeling because of the poor diagnostic quality of items.

## Work in Progress

Today's language learners are surrounded by increasingly complex learning environments. Advanced technologies can provide multiple avenues for observing, assessing, and tracking learners' cognitive progression. The term "learning progression" refers to a hypothetical model of a long-term learning pathway that has been empirically validated (Duschl et al. 2011). These progressions represent a shift in focus from assessment of discrete knowledge to a set of cognitive and metacognitive skills that students need to master across domains. Importantly, it is agreed that there is not necessarily a single "correct" pathway for learning progressions to proceed (Shavelson 2010). Twenty-first century assessment needs to be adaptive enough to observe and track individual students' dynamic progress in cognition, metacognition, and affect (Mislevy et al. 2008). Mislevy et al. argue that the concepts and language of past testing practice are limited in exploring assessment adaptivity. Parshall et al. (2002) explain how computer-assisted assessment can enhance adaptivity through innovations in key dimensions including item format, response action, media inclusion, level of interactivity, scoring, and communication of test results.

In assessing students' speaking proficiency, digitized speech and video offer greater authenticity to assessment, and automated speech recognition techniques allow students to record their voice while performing on a task (Hubbard 2009). For example, automated rating programs, such as Educational Testing Service's Criterion, provide detailed feedback based on error analysis (Chapelle and Douglas 2006). In assessing reading comprehension, computer-assisted assessment programs can gather data about the degree of automaticity in cognitive processing by observing the quantity and quality of readers' attention through eye-tracking methods and

the use of comprehension aids, such as glosses and electronic dictionaries from computer logs.

With technological advances, researchers are currently investigating on the potential of neurophysiological methods, such as eye-tracking and facial expression detection to understand processes involving choice behavior (Rayner 1998; Winke 2013) or language processing (Bax 2013; Conklin and Pellicer-Sánchez 2016). For example, research on eye movement suggests that shorter regressions (reading back in a text) and longer fixations than typical may indicate comprehension difficulties (Rayner 1998). Bax (2013) studied differences in cognitive processing among students with different proficiency levels by examining fixation patterns and search reading patterns. He reported that proficient and less proficient learners show significant differences in terms of lexical and syntactic processing while no evident difference is found for higher cognitive processing. Further, his eye-tracking data revealed that less proficient learners tended to spend more time searching a larger chunk of text whereas more proficient learners were better at expeditious reading by locating key information in the text. This was further substantiated using interview data, suggesting that the observed difference was attributable to metacognitive strategies. Bax noted that the nature of cognitive processing involved in test-taking situations depends on the quality of test items, that is, the extent to which they elicit a wide range of cognitive processes while distinguishing among students with different proficiency levels. Some of the items that he examined failed to elicit distinguishable cognitive processes. This may indicate that multiple processing strategies co-occurred, which is difficult to capture through the eye-tracking method.

Brunfaut and McCray (2015) examined cognitive processes of Aptis reading tasks using eye-tracking and stimulated recall methods. They found that high-proficient learners spend less time fixating on the task prompts and response options, indicating higher efficiency in information processing. While the study participants reported a wide range of processing skills (lexical processing > creating propositional meaning > inferencing > syntactic parsing > creating a text-level representation), they tended to adopt careful reading at both local and global levels as defined by Khalifa and Weir (2009). Expeditious reading using skimming and search strategies were less frequently observed. The authors noted that observed differences in processing were associated with task type more than proficiency levels targeted by the tasks. For example, the "gap-fill" task type elicited lower-level processing involving careful local reading, whereas "sentence-ordering and matching headings" types invoked higher-level processing involving careful global reading and expeditious reading to a less degree. These findings are consistent with findings from previous studies reviewed earlier in this chapter. These two studies illustrate how triangulating data from verbal reports with neurophysiological eye movements allows for more thorough insights into both low- and high-level processing with different reading strategies.

Along with eye-tracking methods, technological advances in facial expression analysis may be used to gather information about students' emotion, intention, cognitive processing, physical effort, and pragmatics synchronously as they engage in assessment tasks. Corpus-based computer systems are currently developed in

order to automatically analyze and recognize facial motions and expressions from visual data (Fasel and Luettin 2003). They are being applied to many areas such as paralinguistic communication, clinical and cognitive psychology, neuroscience, pain analysis, and multimodal human computer interfaces. Yet, their applications to language learning and assessment are unknown to the best of my knowledge. In the near future, these technological influences will have a profound impact on what to assess and how to assess it. It will allow for the assessment of learners' cognitive progression through a dynamic interplay with other traits in digital environments.

## Problems and Difficulties

### Factor Analytic Approach to Skill Identification

Methodological limitations are the greatest challenge in assessing learner cognition and its dynamic interaction with other intra- and interpersonal factors. Conceptually, a cognitive skill is idiosyncratic and context dependent. It is dynamical instead of static because it constantly interacts with other similar skills as well as other related factors. Traditional psychometric approaches often fail to identify multiple skills separately. Common factor analytic approaches are inappropriate for identifying highly correlated skills because factor loadings are essentially based on the contribution of items to test performance (Weir and Khalifa 2008). Furthermore, linear factor analysis of the observed item-pair correlation matrix introduces confounding factors associated with difficulty levels partly due to its linearity assumption in violation of the nonlinear nature of the item response functions (Jang and Roussos 2007). The cognitive diagnostic modeling approaches reviewed briefly in this chapter can handle the cognitive interaction between skills and test items. However, most diagnostic models presuppose strong assumptions about the inter-skill relationships (Jang 2009a). More importantly, valid diagnostic inferences from such modeling depend on whether students' performance data come from a cognitively engineered assessment for the purpose of skill diagnosis. This issue cannot be fully addressed by fit statistics because it is a theoretical matter that requires substantive expertise and evidentiary triangulation.

### Efficacy of Strategy Use

Another methodological issue is that determining skill proficiency based on the observed frequencies of strategy use is problematic. Many different scenarios of strategy use are possible and not all of them lead to successful performance as discussed earlier in this chapter. Depending on task type and diagnostic quality, multiple skills may be required for the same task but be prioritized differently by test takers with different proficiency and background. Alternatively, they may be processed in a compensatory manner. For example, when a task requires the

application of multiple skills involving both grammatical knowledge and higher-level processes conjunctively, successful test takers typically integrate these skills, whereas unsuccessful test takers may overly rely on higher-level skills to compensate for a lack of vocabulary or syntactic knowledge (Brunfaut and McCray 2015; Harding et al. 2015).

A proficient language user is a good strategy user who actively seeks to construct and communicate meaning through the strategic use of various linguistic and sociolinguistic resources at different levels of mental processes. However, identifying and classifying learner strategies have been subject to debate because of difficulty differentiating construct-relevant from construct-irrelevant strategies and the interdependence of strategy categories (Cohen 2012). One recurring issue is that content experts do not agree on what skills are targeted by test items (Alderson 2005). It is possible that content experts may use different grain sizes in identifying skills. Both the number and type of skills can vary significantly depending on the grain size of skill that content experts refer to, especially when item content analysis is mainly used without data informing students' actual mental processes. While determining an ideal grain size needs to be based on empirically validated relevant theories, it can be guided by intended pedagogical effects as well. Testing too many skills with not enough items per skill will not only jeopardize the measurement precision but also make it difficult for users to utilize resulting diagnostic information for planning future actions. Therefore, determining what skill and how many skills to be tested should consider both process data and practical implications of the skills-based assessment for learning.

Another issue lies in the fact that low-proficient learners do not lack strategies. Instead, their problems are more likely to do with the inappropriate use of strategies (Abraham and Vann 1987). This becomes more prominent with intermediate and advanced learners who can self-regulate their strategy use effectively. Research tends to focus on the frequency of strategies without considering their effect on performance and the degree of self-regulation accompanied with strategy use. Although difficulty in differentiating good from poor learners may be explained in terms of the type of strategy used, it is often difficult to determine the efficacy of strategy use due to individual and contextual differences.

This context-dependent nature of cognitive processing helps emphasize the importance of context in understanding how the human mind works differently across different sociocultural settings. It calls for an alternative view of how the mind works in language learning and innovative assessment that "takes into account" interactions with physical, social, and cultural environments instead of statistically controlling for such influence. Different cognitive processes and associated strategies may be prioritized across different cultural and learning contexts. Much research in second language acquisition has been devoted to text analysis in order to understand ways in which different languages shape the mind, known as the Sapir-Whorf hypothesis (Connor 1996). However, the relationship between cognition and culture has not been fully appreciated in the field of language assessment because any cultural influence is considered a threat to construct validity in high-stakes testing.

## Future Directions

As there is a growing interest in learning-oriented assessment approaches, the field of language assessment should embrace an alternative view of and methodological approach to assessing test takers' cognitive potential while capitalizing on the contextual specificity. According to complex dynamic systems theory, a learner is a dynamic system with many components interacting with each other (Smith and Thelen 2003; van Geert and van Dijk 2002). It is through the interaction of these intrapersonal variables and the influence of environmental factors that the learner's mind can be better understood. Although intraindividual variability is previously considered to be a result of measurement error interpreted as "noise," a dynamic systems perspective sees it as a "driving force of development" (van Geert and van Dijk 2002, p. 4). Future language assessment should provide the opportunity to assess, model, track, and scaffold progressions as a result of such interactions (Jang et al. 2015). In this way, we may be able to catch a glimpse of such complex and dynamical learner cognition.

We may continue to measure language ability in the standard way, but information from such measurements has few implications for both advancing theories and making positive impact on the lives of people involved. In doing so, we will continue to neglect the importance of metacognitive and motivational consciousness in learning, emotions involved in thinking and as a consequence of actions, and the broader social context in which we are embedded. In fact, few would argue against the need to assess learner cognition in conjunction with its interaction with internal and external factors. In my view, what matters now is whether or not we are equipped with the necessary methodological innovations. I strongly support technological advances that allow us to make observations (and data gathering) less intrusive and less dependent on self-reported data. For example, data mining and machine learning as approaches to assessing learner cognition and growth during language learning may help us innovate current assessment approaches. Multichannel data involving physiological, neural, and behavioral traces may provide information that can be used to complement self-reported data. Data gathering and analysis in the real time while students perform on a task will help assessment provide adaptive scaffolding for individual students. All of these approaches must be grounded in well-reasoned student models of cognition, be supported by evidentiary reasoning, and be engaged in while bearing potential effects in mind. The pendulum swings both ways. We may continue to theorize learner cognition without practical implications or attempt to program the mind using neural networks without theories. Neither will advance the field of language assessment nor shall we.

## Cross-References

▶ Assessing Meaning
▶ Assessing Second/Additional Language of Diverse Populations
▶ Dynamic Assessment

## Related Articles in the Encyclopedia of Language and Education

James Cummins: BICS and CALP: Empirical and Theoretical Status of the Distinction. In Volume: Literacies and Language Education

Nich C. Ellis: Implicit and Explicit Knowledge About Language. In Volume: Language Awareness and Multilingualism

Claudia Finkbeiner, Joanna White: Language Awareness and Multilingualism: A Historical Overview. In Volume: Language Awareness and Multilingualism

Marjolijn H. Verspoor: Cognitive Linguistics and Its Applications to Second Lan guage Teaching. In Volume: Language Awareness and Multilingualism

## References


Abraham, R. G., & Vann, R. J. (1987). Strategies of two language learners: A case study. In A. L. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 85–102). London: Prentice Hall International.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Percsich, R., & Szabo, G. (2000). Sequencing as an item type. *Language Testing, 17*(4), 423–447.

Anderson, J. R. (2000). *Cognitive psychology and its implication* (5th ed.). New York: Worth.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. *Handbook of reading research, 1*, 255–291.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441–465.

Brunfaut, T., & McCray, G. (2015). Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study. *ARAGs Research Reports Online,* Vol. AR/2015/001. London: The British Council.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Canale, M., & Swain, M. (1980). Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Carrell, P. L. (1988). SLA and classroom instruction: Reading. *Annual Review of Applied Linguistics, 9*, 223–242.

Carrell, P. L. (1998). Can reading strategies be successfully taught?. *Australian Review of Applied Linguistics, 21*, 1–20.

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and individual differences, 19*(2), 246–251.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology* (Cambridge language assessment series). Cambridge: Cambridge University Press.

Cohen, A. D. (2012). Test taker strategies and task design. In G. Fulcher & F. Davison (Eds.), *The Routledge handbook of language testing* (pp. 262–277). Abingdon: Routledge.

Cohen, A. D., & Upton, T. A. (2007). "I want to go back to the text": Response strategies on the reading subtests of the New TOEFL. *Language Testing, 24*(2), 209–250.

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 0267658316637401.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second language writing*. Cambridge: Cambridge University Press.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education, 47*(2), 123–182.

Ehrlich, M. F., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing, 11*, 29–63.

Ericsson, K. A. (2003). The acquisition of expert performance as problem solving: Construction and modification of mediating mechanisms through deliberate practice. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 31–83). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed., 1984). Cambridge, MA: The MIT Press.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recognition, 36*(1), 259–275.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist, 6*(1), 126–135.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317–336.

Hruby, G. G., & Goswami, U. (2011). Neuroscience and reading: A review for reading education researchers. *Reading Research Quarterly, 46*(2), 156–172.

Hubbard, P. (2009). *Computer assisted language learning: Critical concepts in linguistics* (Vol. I-IV). London/New York: Routledge.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished PhD dissertation, University of Illinois at Urbana Champaign, Urbana.

Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

Jang, E. E. (2009b). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*, 210–238.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance–based nonparametric approach. *Journal of Educational Measurement, 44*(1), 1–21.

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. (2015). Mediation of goal orientations and perceived ability on junior students' responses to diagnostic feedback. *Language Testing, 32*(3), 299–316.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL).* Unpublished doctoral dissertation, University of Illinois, Urbana Champaign.

Kendeou, P., Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research and Practice, 29*(1), 10–16.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge, UK: Cambridge University Press.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced Reparameterized Unified Model. *Language Testing, 28*(4), 509–541.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Koda, K. (2005). *Insights into second language reading*. New York: Cambridge University Press.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R., Chapelle, C. A., Chung, Y.-R., & Xu, J. (2008). Options for adaptivity in computer-assisted language learning and assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 9–24). Ames: Iowa State University.

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575–603.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist, 37*(2), 91–105.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Purpura, J. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 47*(2), 289–325.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Shavelson, R. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford: Stanford University Press.

Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*(4), 513–521.

Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*(8), 343–348.

Spolsky, B. (1990). Social aspects of individual assessment. In J. de Jong & D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 3–15). Avon: Multilingual Matters.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32–71.

Stanovich, R. J. (2001). Metacognition, abilities, and developing expertise: What makes an expert student? In H. J. Hartman (Ed.), *Metacognition in learning and instruction: Theory, research, and practice* (pp. 247–260). Dordrecht: Kluwer.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior & Development, 25*, 340–374.

Weir, C. J. (2005). *Language testing and validation: An evidenced-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J., & Khalifa, H. (2008). *A cognitive processing approach towards defining reading comprehension. Research Notes,* UCLES 2008. Cambridge, UK: Cambridge English Language Assessment.

Winke, P. (2013). Eye-tracking technology for reading. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1029–1046). Hoboken: Wiley-Blackwel.

# Criteria for Evaluating Language Quality

Glenn Fulcher

**Abstract**

The assessment of language quality in the modern period can be traced directly to the work of George Fisher in the early nineteenth century. The establishment of a scale with benchmark samples and tasks has been replicated through Thorndike (1912) and into the present day. The tension between assessing observable attributes in performance and underlying constructs that makes performance possible is as real today as in the past. The debate impacts upon the way scales and descriptors are produced, and the criteria selected to make judgments about what constitutes a quality performance, whether in speech or writing. The tensions work themselves through the history of practice, and today we find ourselves in a pluralistic philosophical environment in which consensus has largely broken down. We therefore face a challenging environment in which to address the pressing questions of evaluating language quality.

**Keywords**

Language quality • Performance assessment • Rating scales • Descriptors • Rating criteria

G. Fulcher (✉)
English Department, School of Arts, University of Leicester, Leicester, UK
e-mail: gf39@le.ac.uk

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_13

## Contents



## Introduction

The Oxford English dictionary defines "quality" as "the standard of something as measured against other things of a similar kind; the degree of excellence of something." In language testing the "something" is a language product, which may be a sample of talk or writing. This is "measured" against similar products that have been independently assessed as being appropriate for a particular communicative purpose. The quality of the language sample is the window into the ability of its producer. Or as Latham puts it:

> . . .we cannot lay bare the intellectual mechanism and judge of it by inspection, we can only infer the excellence of the internal apparatus and the perfection of its workmanship from the quality of the work turned out. (Latham 1877, p. 155)

The first attempt to measure language quality by comparison with other samples is found in Fisher's Scale Book (Fulcher 2015a). Between 1834 and 1836, while headmaster of the Royal Hospital School in London, Fisher developed his scale book, in which language performance was classified into five major levels, each with quarter intervals. This produced a 20-level scale. Each level was characterized by writing samples that represented what a pupil was expected to achieve at that level. For spelling there were word lists, and for speaking there were lists of prompts/tasks that should be undertaken successfully. The Scale Book has not survived, but it is clear that Fisher had invented a method for the measurement of quality that is still in use today. There is clear evidence that Thorndike had seen, or was aware of, Fisher's methods (Fulcher 2015b, pp. 84–88). With reference to the assessment of French and German, he suggested attaching performance samples to

levels, together with a brief description of what could be achieved at each level (Thorndike 1912).

It is not clear what criteria were used by Fisher or Thorndike for the selection of samples to characterize each level, other than the professional judgment of experts familiar with the context of the use of the scale. For Fisher, this was a school context in which boys were being educated in preparation for a life in the navy. Thorndike also had a US high school context in mind, but his focus was psychometric and methodological, rather than practical hands-on assessment. But what is clear in both cases is that new language samples collected in an assessment are being evaluated in comparison with criterion samples. Although the term would not be invented for many decades, Fisher was the first to employ criterion-referenced assessment in an educational context.

## Early Developments

The use of criteria external to the assessment context has been central to the evaluation of language quality from the start. It is important to remember that the "criteria" of "criterion-referenced" assessment are not abstract levels that today are frequently referred to as "standards." Rather, the term "criterion" and "standard" were used interchangeably to refer to real-world behaviors that a test taker would be expected to achieve in a non-test environment (Glaser 1963; Fulcher and Svalberg 2013). In the development of the first large-scale language test during the First World War, it was therefore considered essential to reflect such real-world behavior in test content (Fulcher 2012). A group and an individual test of English as a second language were developed to identify soldiers who should be sent to language development batteries rather than deployed to active service. Yerkes (1921, p. 335) reports that the individual test was to be preferred because it was possible to make the content reflect military language more than the group test. Of course, the tasks were still a considerable abstraction from real life, but the criterion was nevertheless the kind of language that was contained in "the drill" (see Fulcher 2015b, pp. 135–140). The score on the test items was interpreted by matching it to a level descriptor from A to E that provided score meaning in absolute criterion terms:

> Men can be tested for English-speaking ability and rated on a scale of A, B, C, D, E. In language the rating E means inability to obey the very simplest commands unless they are repeated and accompanied by gestures, or to answer the simplest questions about name, work, and home unless the questions are repeated and varied. Rating D means an ability to obey very simple comments (e.g., "Sit down," "Put your hat on the table"), or to reply to very simple questions without the aid of gesture or the need of repetition. Rating C is the level required for simple explanation of drill; rating B is the level of understanding of most of the phrases in the Infantry Drill Regulations; rating A is a very superior level. Men rating D or E in language ability should be classified as non-English. (Yerkes 1921, p. 357)

From the First World War, assessing the quality of language performances had two critical components. First was the explicit criterion-referenced relationship

between the content of the test and the domain to which prediction was sought. Second is the level descriptor that summarized what a test taker at a particular level could do with the language in the non-test domain. These two components of performance tests allowed numerical scores to be invested with real-world meaning.

The interwar period was marked by the massive expansion of state provided education in many Western countries. Assessment became critical for accountability, and accountability required controlling the costs of assessment in large systems. There was therefore a focus on the "new-type" multiple choice tests at the expense of performance (Wood 1928). When a new need to assess language performance reemerged in the Second World War, it was as if everything that had been learned during the First World War needed to be reinvented. Thus it was that Kaulfers and others working in the Army Specialized Training Program (ASTP) had to develop new performance tests and descriptors:

> The nature of the individual test items should be such as to provide specific, recognisable evidence of the examinee's readiness to perform in a life-situation, where lack of ability to understand and speak extemporaneously might be a serious handicap to safety and comfort, or to the effective execution of military responsibilities. (Kaulfers 1944, p. 137)

It is the criterion-referenced nature of the decisions being made that requires the quality of language to be assessed through performance. The touchstone was learning to speak a colloquial form of a second language, rather than learning *about* the language (Agard and Dunkel 1948; Velleman 2008). Unlike the individual test created by Yerkes in 1917, the tasks were not domain specific to the military, but covered the functions of securing services and asking for and giving information. This was all that could be achieved in the 5 mins allocated to an individual test. Kaulfers reports that language quality was assessed according to the two criteria of *scope* and *quality* of speech:

**Scope of Oral Performance**

(a) Can make known only a few essential wants in set of phrases of sentences.
(b) Can give and secure the routine information required in independent travel abroad.
(c) Can discuss the common topics and interests of daily life extemporaneously.
(d) Can converse extemporaneously on any topic within the range of his knowledge or experience.

**Quality of Oral Performance**

(0) Unintelligible or no response. A literate native would not understand what the speaker is saying, or would be confused or mislead.
(1) Partially intelligible. A literate native might be able to guess what the speaker is trying to say. The response is either incomplete, or exceedingly hard to understand because of poor pronunciation or usage.
(2) Intelligible but labored. A literate native would understand what the speaker is saying, but would be conscious of his efforts in speaking the language. The delivery is hesitating, or regressive, but does not contain amusing or misleading errors in pronunciation or usage.
(3) Readily intelligible. A literate native would readily understand what the speaker is saying, and would not be able to identify the speaker's particular foreign nationality. (Kaulfers 1944, p. 144)

Under "quality" we can see the emergence of two themes that remain issues of research and controversy to this day. The first is the nature of "intelligibility" and its relation to "comprehensibility," given the constant reference to pronunciation (see Browne and Fulcher 2017). Second is the reference to a "literate" (later to be termed "educated") native speaker as the intended interlocutor.

The ASTP program scored language quality at three levels, under the four headings of fluency, vocabulary, pronunciation and enunciation, and grammatical correctness. The scale for fluency shows that the metaphorical nature of the construct as "flowing" like a river (Kaponen and Riggenbach 2000) emerged very early in performance assessment:

> Fluency
> (2) Speaks smoothly, phrasing naturally according to his thoughts.
> (1) Occasionally hesitates in order to search for the right word or to correct an error.
> (0) Speaks so haltingly that it is difficult to understand the thought he is conveying.
> (Agard and Dunkel 1948, p. 58).

Qualitative level descriptors that closely resemble these early examples have been used ever since, even if they have frequently been disassociated with their original criterion-referenced meaning. They are normally placed in a *rating scale* consisting of two or more levels. Language samples or tasks that are claimed to typify a particular level may be used, following the early practices of Fisher and Thorndike. The rating scale is normally used to match a performance with the most relevant description to generate a score.

## Major Contributions

It should not be surprising that some of the most important contributions have been made within the military context. After the Second World War, the Foreign Service Institute (FSI) was established in the United States to forward the wartime assessment agenda. Although it is still frequently claimed that what emerged from the US military as the Foreign Service Institute (FSI) rating scale was decontextualized (devoid of context, content, or performance conditions) (Hudson 2005, p. 209), as early as 1958 descriptors were attached to the FSI scale. The following example illustrates the level of contextualization that was present:

> FSI Level 2: Limited Working Proficiency.
> Able to satisfy routine social demands and limited work requirements.
> Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e., topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite

> accurately but does not have thorough or confident control of the grammar. (reproduced in Fulcher 2003, p. 226)

The level of contextualization is problematic. The wording suggests "tasks" that a speaker might successfully undertake and the quality of language that might be produced. Yet, it is not specific to its intended military purpose. This is an issue which still exercises language testers today. On the one hand is the argument that all descriptors and scales should refer to constructs only and avoid any reference to context (Bachman and Savignon 1986). The primary purpose of non-contextualization is to achieve greater generalizability of scores across test tasks and real-world contexts. What the language tester is interested in is the underlying constructs or abilities that make communication possible. On the other hand is the argument that by limiting score meaning to specified domains, validation becomes an achievable goal.

The halfway house of the FSI has survived to the present day, despite debates for and against domain specificity. During the 1960s the language and format of the FSI descriptors became standard throughout the military and security agencies in the United States, resulting in a description of language performance known as the Interagency Language Roundtable (ILR), which is still in use today (Lowe 1987). The ILR has also formed the basis of the North Atlantic Treaty Organization (NATO) approach to scoring language quality (Vadász 2012), articulated in Standardization Agreement 6001 (STANAG 6001). The language and structure of the descriptors follows the ILR closely, although additional references to topics and functions have been added in its various revisions (NATO 2010).

The assessment of language quality in the military soon spread to the educational sector. In the early 1980s the American Council on the Teaching of Foreign Languages (ACTFL) and Educational Testing Service (ETS) received US federal grants to adapt the FSI and ILR to create a description of language performance for wider use. The ACTFL *Guidelines* were published in 1986 and revised in 1999 and 2012 (http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012). They have become the de facto framework for describing language performance in the United States in both education and the workplace (Swender 2003).

These descriptions combine linguistic and nonlinguistic criteria and are assumed to be relevant to all languages. The sequence of descriptors on the scale represents an intuitive understanding of the order of second language acquisition and the increasing complexity of real-world tasks that learners can perform, but for which there is little empirical research evidence (Brindley 1998; Chalhoub-Deville and Fulcher 2003).

The FSI descriptors and their subsequent use both in the military and educational sectors have had a profound impact upon the structure and wording of all subsequent scales used for evaluating language quality. The theoretical assumptions, and even the wording, can be traced in all extant scales.

While ACTFL is the dominant system in the United States, the Canadian Language Benchmarks (CLB) is used in Canada, and the Common European Framework of Reference (CEFR) has been developed for use in Europe. These are institutionalized systems and therefore have had a very wide impact on practice (Liskin-Gasparro 2003). However, while both the ACTFL and CLB were designed for operational use in rating language performance, the CEFR bears the hallmarks of a set of abstract standards that cannot be simply taken and used in real assessment contexts (Jones and Saville 2009).

The CLB was developed to assess the English of adult immigrants to Canada and is "a descriptive scale of communicative proficiency in English as a Second Language (ESL) expressed as 12 benchmarks or reference points" (Pawlikowska-Smith 2000, p. 7). Pawlikowska-Smith (2002) argues that the CLB is based on a model of communicative proficiency, drawing specifically on notions of linguistic, textual, functional, sociocultural, and strategic competence, adapted from Bachman and Palmer (1996) and Celce-Murcia et al. (1995). There are three general levels (basic, intermediate, and advanced), each with four subdivisions, for each of the four skill competencies (speaking, listening, reading, and writing).

The CEFR aims to be a pan-European framework for teaching and testing languages (Council of Europe 2001). Like the CLB it has three general levels of basic, independent, and proficient, each subdivided into two levels, providing a six-level system. The system comprises two parts. The first is a qualitative description of each level. For speaking and writing it is elaborated in productive, receptive, and interactive modes. This is "horizontal" in that it does not attempt to help distinguish between levels; it is a taxonomy of the things that language learning is about. The second part is a quantitative description of the levels in terms of "can-do" statements. This is "vertical" in that the levels are defined in terms of hierarchical descriptors.

## Work in Progress

### Rating Scale Development

The major contributions are all institutional systems that perform a policy role within high-stakes testing systems. They are all intuitively developed scales, with the exception of the CEFR, which is a patchwork quilt of descriptors taken from other scales, constructed using a measurement model based on teacher perceptions of descriptor difficulty (Fulcher 2003, pp. 88–113). Dissatisfaction with linear scales that are unlikely to reflect either processes in SLA or performance in specific domains has led to research in scale development that is "data driven." One approach has been through the application of binary choices to separate writing or speaking samples using critical criteria (Upshur and Turner 1995), which has subsequently been applied to TOEFL iBT (Poonpon 2010). The other main approach is the description of performance data to populate descriptors, whether this be taken from test taker performance on tasks (Fulcher 1996) or

expert performance in real-world contexts through performance decision trees (Fulcher et al. 2011). The goal of the latter enterprise is to create a "thick description" of domain-specific performance, thus establishing a true criterion-referenced description against which to match test-generated performances. Data-driven approaches are also being used in prototype writing scales (Knoch 2011). The selection of scale type for particular assessment contexts is a key issue for current research.

### Construct Definition and Validation

While our understanding of what constitutes reasonable performance in specific domains has increased immensely in recent years, the definition and assessment of particular constructs or abilities that enable such performance has been more problematic. Ongoing research into "interactive competence" is particularly important because of the potential to assess individuals in relation to how their own performance and competence is impacted by others (Chalhoub-Deville 2003). Recent work on interactive patterns (Galaczi 2008) and communicative strategies (e.g., May 2011) represents the ongoing attempt to produce operational assessments with richer interactive construct definitions.

### Test Taker Characteristics

Closely related to how participants interact is the question of how individual characteristics affect interaction. The practical implications of this research may impact on how test takers are selected for pair or group speaking tests. Berry (2007) summarizes her extensive research into the impact of personality type, showing that levels of introversion and extroversion can influence speaking scores. Ockey (2009) also found that assertive test takers score higher in group tests, but that less assertive students were not impacted by the pairing. The differences in findings may suggest that the results are conditioned by cultural factors that require further investigation. Nakatsuhara (2011) has also shown that there is variation by proficiency level, personality, and group size. There is clearly much more work to be done here to identify significant variables and their impact on performance.

## Problems and Difficulties

### Generalizability Versus Specificity

Resistance to the use of data-driven or domain-specific scales in large-scale testing is related to restrictions on score meaning. The underlying issue is what constitutes a "criterion" in criterion-referenced testing. For those who argue that domain-specific

inferencing is paramount, the criterion is the language used in real-world applications, which echoes the "job description" tradition of validation (Fulcher and Svalberg 2013). The claim to generalizability of scores to multiple domains and purposes reverts to a criterion-related validation claim based on correlation with an external measure or comparison group (Fulcher 2015b, pp. 100–102). The tension is between substantive language-based interpretations and psychometric expediency. The latter is sometimes used to advocate a robust financial model of "off-the-peg" test use by testing agencies without the need to provide additional validation evidence for changes in test purpose (Fulcher and Davidson 2009). The interplay between the meaning of "criterion" and the economics of global language test use in policy provides plenty of opportunity for conflict.

## Construct Definition

The new interest in content validation (Lissitz and Samuelsen 2007) combined with a lack of interest in construct language within argument-based approaches to validity (Kane 2012, p. 67) has had an impact on validation practices. Chapelle et al. (2010, pp. 3–4) apply this to the scoring of language samples collected in the TOEFL iBT, which moves directly from observation to score, without the requirement for any intervening construct. The point of debate has therefore moved away from construct definition to whether simple content comparison between test tasks and the domain constitutes validation evidence. Kane (2009, pp. 52–58) argues that validation activity remains with the interpretation of scores, and so while the focus may shift to observable attributes in specific performances, there remains a requirement to demonstrate generalizability to all possible test tasks and extrapolation to a domain that cannot be fully represented. But in the simple content validity stance, and the more complex argument-based stance, the room for generalizability of score meaning is considerably reduced. To what extent should construct language be retained? And just how generalizable are the claims that we can reasonably "validate"?

## Rating Scales Versus Standards

The critique of "frameworks" or "standards" documents as tools for policy implementation (Fulcher 2004) has resulted in a recognition that institutional "scales" cannot be used directly to evaluate language quality (Weir 2005; Jones and Saville 2009; Harsch and Guido 2012). But the power of such documents for the control of educational systems has increased the tendency for misuse (Read 2014). The confusion between "standards" and "assessments" is part of the subversion of validity that has been a by-product of the use of scales to create the equivalent of standardized weights and measures in education, similar to those in commerce (Fulcher 2016). This inevitably draws language testers into the field of political action, even if they take the view that they are merely "technicians" producing tools for decision-making processes.

## Future Directions

What has been achieved in the last decade is quite substantial. When the TOEFL Speaking Framework (Butler et al. 2000) is replaced, the new volume will reflect the very significant progress that has been made in assessing the quality of spoken language. We now have considerably more options for scoring models than the simple "more than. . ..less then. . .." descriptors that characterized rating scales in use since the Second World War. These are likely to be richer because of the advances in domain description and referencing. Our deeper understanding of interaction now also informs task design not only for pair and group assessment but also for simulated conversation in a computer-mediated test environment. These developments will inform critical research in the coming years.

## Domains of Inference

The issue of what is "specific" to a domain has come back to the fore in language testing (Krekeler 2006) through the renewed interest in content and the instrumentalism of argument-based approaches to validation. The emphasis must now be on the understanding of what constitutes successful language use in specific domains. Work in the academic domain to support task design in the TOEFL iBT is noteworthy (Biber 2006), as is work on service encounter interactions (Fulcher et al. 2011). There is a long tradition of job-related domain analysis in applied linguistics (e.g., Bhatia 1993), and language testing practice needs to formulate theory and practices for the inclusion of such research into test design.

## Research on Scoring Instruments

Directly related to the previous issue is research into different types of scoring instruments. The efficacy of task-dependent and task-independent rating scales depending on test purpose requires further investigation (Chalhoub-Deville 1995; Hudson 2005; Jacoby and McNamara 1999; Fulcher et al. 2011). As we have found it more difficult to apply general scales to specific instances of language use, it becomes more pressing to show that descriptors adequately characterize the performances actually encountered and can be used reliably by raters (Deygers and Van Gorp 2015).

## Policy Analysis

The most influential approaches to describing language quality are those with the support of governments or cross-border institutions, where there is great pressure for systems to become institutionalized. The dangers associated with this have been

outlined (Fulcher 2004; McNamara 2011), but the motivations for the institutionalization of "frameworks" need further investigation at level of policy and social impact. Of particular concern is the need of bureaucrats to create or defend regional identities or language economies. This leads to the danger that the language testing industry makes claims for tests that cannot be defended and may be particularly dangerous to individual freedoms. As Figueras et al. (2005, pp. 276–277) note, "linkage to the CEFR may in some contexts be required and thus deemed to have taken place. . .."

## Living with Plurality

The immediate post-Messick consensus in educational assessment and language testing has broken down (Newton and Shaw 2014). Fulcher (2015b, pp. 104–124) discusses four clearly identifiable approaches to validity and validation that have emerged in language testing, some of which are mutually incommensurable. At one end of the cline is the emergence of strong realist claims for constructs resident in the individual test taker, and at the other is an approach to co-constructionism that argues for the creation and dissolution of "constructs" during the act of assessing. This clash of philosophies is not new in language testing, but it is more acute today than it has been in the past. The debate over philosophical stance is probably one of the most important to be had over the coming decade, as it will determine the future epistemologies that we bring to bear on understanding the quality of language samples.

## Cross-References


- ▶ Critical Language Testing
- ▶ History of Language Testing
- ▶ Methods of Test Validation
- ▶ Qualitative Methods of Validation
- ▶ The Common European Framework of Reference (CEFR)


## Related Articles in the Encyclopedia of Language and Education


Olga Kagan, Kathleen Dillon: Issues in Heritage Language Learning in the United States. In Volume: Second and Foreign Language Education

Sandra Lee McKay: Sociolinguistics and Language Education. In Volume: Second and Foreign Language Education

Amy Ohta: Sociocultural Theory and Second/Foreign Language Education. In Volume: Second and Foreign Language Education

## References


Agard, F., & Dunkel, H. (1948). *An investigation of second language teaching*. Chicago: Ginn and Company.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*(4), 380–390.

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Harlow: Longman.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In P. Trofimovich & T. Isaacs (Eds.), *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives* (pp. 37–53). London: Multilingual Matters. https://zenodo.org/record/165465#.WDltUbTfWhD.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper*. Princeton: Educational Testing Service.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5–35.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals, 36*(4), 409–506.

Chapelle, C. A., Enright, M., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Council of Europe. (2001). *Common European Framework of reference for language learning and teaching*. Cambridge: Cambridge University Press.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, Online First* April 7, 1–21. doi:10.1177/0265532215575626.

Figueras, N., North, B., Takala, S., Van Avermaet, P., & Verhelst, N. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*(3), 261–279.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253–266.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). London/New York: Routledge.

Fulcher, G. (2015a). Assessing second language speaking. *Language Teaching, 48*(2), 198–216.

Fulcher, G. (2015b). *Re-examining language testing: A philosophical and social inquiry*. London/New York: Routledge.

Fulcher, G. (2016). Standards and frameworks. In J. Banerjee & D. Tsagari (Eds.), *Handbook of second language assessment*. Berlin: DeGruyter Mouton.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123–144.

Fulcher, G., & Svalberg, A. M.-L. (2013). Limited aspects of reality: Frames of reference in language assessment. *International Journal of English Studies, 13*(2), 1–19.
Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.
Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119.
Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.
Harsch, C., & Guido, M. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*(4), 228–250.
Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25*, 205–227.
Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241.
Jones, N., & Saville, N. (2009). European language policy: Assessment, learning and the CEFR. *Annual Review of Applied Linguistics, 29*, 51–63.
Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Charlotte: Information Age Publishing.
Kane, M. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 66–70.
Kaponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). Ann Arbor: University of Michigan Press.
Kaulfers, W. V. (1944). War-time developments in modern language achievement tests. *Modern Language Journal, 70*(4), 366–372.
Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96.
Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing, 23*(1), 99–130.
Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Dighton, Bell and Company.
Liskin-Gasparro, J. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*(3), 483–490.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437–448.
Lowe, P. (1987). Interagency language roundtable proficiency interview. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 43–46). Washington, DC: TESOL Publications.
May, L. A. (2010). Developing speaking assessment tasks to reflect the 'social turn' in language testing. *University of Sydney Papers in TESOL, 5*, 1–30.
May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127–145.
McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–349.
McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(04), 500–515.
Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483–508.
NATO. (2010). *STANAG 6001 NTG language proficiency levels* (4th ed.). Brussels: NATO Standardization Agency.
Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. London: Sage.

Ockey, G. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161–186.

Pawlikowska-Smith, G. (2000). *Canadian language benchmarks 2000: English as a second language – For adults*. Toronto: Centre for Canadian Language Benchmarks.

Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: Theoretical framework*. Toronto: Centre for Canadian Language Benchmarks.

Poonpon, K. (2010). Expanding a second language speaking rating scale for instructional assessment purposes. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 8*, 69–94.

Read, J. (2014). The influence of the Common European Framework of reference (CEFR) in the Asia-Pacific Region. *LEARN Journal: Language Education and Acquisition Research Network,* 33–39.

Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. *Foreign Language Annals, 36*(4), 520–526.

Thorndike, E. L. (1912). The measurement of educational products. *The School Review, 20*(5), 289–299.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3–12.

Vadász, I. (2012). What's behind the test? *Academic and Applied Research in Military Science, 10*(2), 287–292.

Velleman, B. L. (2008). The "Scientific Linguist" goes to war. The United States A.S.T. program in foreign languages. *Historiographia Linguistica, 35*(3), 385–416.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Wood, B. (1928). *New York experiments with new-type language tests*. New York: Macmillan.

Yerkes, R. M. (1921). *Psychological examining in the United States army* (Memoirs of the National Academy of Sciences, Vol. XV). Washington, DC: GPO.

# Methods of Test Validation

Xiaoming Xi and Yasuyo Sawaki

**Abstract**

Test validation methods are at the heart of language testing research. The way in which validity is conceptualized determines the scope and nature of validity investigations and hence the methods to gather evidence. Validation frameworks specify the process used to prioritize, integrate, and evaluate evidence collected using various methods. This review charts the evolution of validity theory and validation frameworks and provides a brief review of current methodologies for language test validation, organized by the validity inferences to which they are related in an argument-based validation framework. It discusses some problems and challenges associated with our current test validation research and practice and proposes some major areas of research that could help move the field forward.

The argument-based approach to test validation, initially developed for large-scale assessment, will continue to be refined to make it more applicable to test developers and practitioners. Alternative validation approaches for classroom assessment are emerging but could benefit from more empirical verifications to make them theoretically sound as well as practically useful. We are in an exciting era when new conceptualizations of communicative language use such as English as a lingua franca and use of new technologies in real-world communication are pushing the boundaries of the constructs of language assessments. These developments have introduced new conceptual challenges and complexity in redefining the constructs of language assessments and in designing validation research in light of the expanded constructs.

X. Xi (✉)
New Product Development, Educational Testing Service, Princeton, NJ, USA
e-mail: xxi@ets.org

Y. Sawaki
Faculty of Education and Integrated Arts and Sciences, Waseda University, Tokyo, Japan
e-mail: ysawaki@waseda.jp; yasuyo.sawaki@gmail.com

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_14

**Keywords**
Validity • Validation • Validation method • Quantitative methodology • Qualitative methodology

## Contents



## Introduction

Test validation methods are at the heart of language testing research. Validity is a theoretical notion that defines the scope and nature of validation work, whereas validation is the process of developing and evaluating evidence for a proposed score interpretation and use. The way in which validity is conceptualized determines the scope and nature of validity investigations and hence the methods to gather evidence. Validation frameworks specify the process used to prioritize, integrate, and evaluate evidence collected using various methods. Therefore, this review will delineate the evolution of validity theory and validation frameworks and synthesize the methodologies for language test validation.

In general, developments of validity theories and validation frameworks in language testing have paralleled advances in educational measurement (Cureton 1951; Cronbach and Meehl 1955; Messick 1989; Kane 1992). Validation methods have been influenced by three areas in particular. Developments in psychometric and statistical methods in education have featured prominently in language testing research (Bachman and Eignor 1997; Bachman 2004). Qualitative methods in language testing (Banerjee and Luoma 1997) have been well informed by second language acquisition, conversation analysis, and discourse analysis. Research in cognitive psychology has also found its way into core language testing research, especially that regarding introspective methodologies (Green 1997) and the influence of cognitive demands of tasks on task complexity and difficulty (Iwashita et al. 2001).

## Early Developments

Earlier conceptualizations of validity for language tests, represented by Lado and Clark (Lado 1961; Clark 1978), focused on a few limited *types* of validity, such as content validity and predictive or concurrent validity, which support primarily score-based predictions, rather than theoretically and empirically grounded explanations of scores that provide the basis for predictions. Because validity was treated as consisting of different types instead of a unitary concept, test providers could conveniently select only one type as sufficient to support a particular test use. Further, test-taking processes and strategies and test consequences were not examined.

In keeping with how validity was conceptualized from the 1950s through late 1970s, the validation methods were limited to correlational analyses and content analyses of test items. Another fairly common line of validation research in the 1960s and 1970s employed factor analytic techniques to test two competing hypotheses about language proficiency.

## Major Contributions

The late 1970s and early 1980s witnessed the first introduction of the notion of construct validity (Cronbach and Meehl 1955) in language testing (Palmer et al. 1981). During the 1980s, there was a shift of focus from predictive or concurrent validity studies to explorations of test-taking processes and factors affecting test performance (Bachman 2000). These studies attested to the growing attention to score interpretation based on empirically grounded explanations of scores.

As validity theories in educational measurement advanced in the 1980s and culminated in Messick's explication of validity (1989), different types of validity became pieces of evidence that supported a *unitary* concept of construct validity, highlighting the importance of combining different types of evidence to support a particular test use. Messick also formally expanded validity to incorporate social values and consequences, arguing that evaluation of social consequences of test use as well as the value implications of test interpretation both "presume" and "contribute to" the construct validity of score meaning (p. 21).

Messick's unitary validity model quickly became influential in language testing through Bachman's work (1990). However, although theoretically elegant, Messick's model is highly abstract and provides practitioners limited guidance on the process of validation.

To make Messick's work more accessible to language testers, Bachman and Palmer (1996) proposed the notion of test usefulness. They discussed five qualities: construct validity, reliability, authenticity, interactiveness, and impact, as well as practicality, which functions to prioritize the investigations of the five qualities. Because of its value in guiding practical work, this framework quickly came to dominate empirical validation research and became the cornerstone for language test development and evaluation. Nevertheless, this formulation does not provide a *logical* mechanism to prioritize the five qualities and to evaluate overall test

usefulness. Since the trade-off of the qualities depends on assessment contexts and purposes, evaluations of overall test usefulness are conveniently at the discretion of test developers and validation researchers.

Following the shift in focus of validity investigations to score interpretation for a particular test use rather than the test itself, theories of validity, impact, ethics, principles of critical language testing (Shohamy 2001), policy and social considerations (McNamara 2006), and fairness (Kunnan 2004) have been formulated to expand the scope of language test quality investigations (Bachman 2005). Although some aspects of their work contribute to the validity of test score interpretations or use, others address broader policy and social issues of testing, which may not be considered as qualities of particular tests (Bachman 2005).

During this period, empirical validation research flourished to address more aspects of validity including factors affecting test performance, generalizability of scores on performance assessments, and ethical issues and consequences of test use (Bachman 2000). Furthermore, applications of sophisticated methodologies, both quantitative (Kunnan 1998) and qualitative (Banerjee and Luoma 1997), became more mature.

## Work in Progress

The search for a validation framework that is theoretically sound but more accessible to practitioners continues. The major development of an argument-based approach to test validation in education measurement (Kane 1992; Kane et al. 1999) has inspired parallel advancements in language testing, represented by Bachman (2005), Bachman and Palmer (2010) and Chapelle et al. (2008).

The notion of a validity argument is nothing new to the field of educational measurement. Nearly two decades ago, Cronbach (1988) started to think of validation as supporting a validity argument through a coherent analysis of all the evidence for and against a proposed score interpretation. Kane and his associates have taken up on this and formalized the development and evaluation of the validity argument by using practical argumentation theories. They see validation as a two-stage process: constructing an interpretive argument, and developing and evaluating a validity argument. They propose that for each intended use of a test, an interpretive argument is articulated through a logical analysis of the chain of inferences linking test performance to a decision and the assumptions upon which they rest. The assumptions, if proven true, lend support for the pertinent inference. The network of inferences, if supported, attaches more and more meaning to a sample of test performance and the corresponding score so that a score-based decision is justified. The plausibility of the interpretive argument is evaluated within a validity argument using theoretical and empirical evidence. Their approach also allows for a systematic way to consider potential threats to the assumptions and the inferences and allocate resources to collect evidence to discount or reduce them.

This conceptualization has not expanded the scope of validity investigations beyond that of Messick (1989), which provides the most comprehensive and in-depth discussion of values of score interpretations and consequences of test uses (McNamara 2006). However, the major strength of Kane's approach lies in providing a transparent working framework to guide practitioners in three areas: prioritizing different lines of evidence, synthesizing them to evaluate the strength of a validity argument, and gauging the progress of the validation efforts. It has considerable worth in helping answer three key questions: where to start, how strong the combined evidence is, and when to stop.

Although test use and consequences were omitted in the earlier developments of his framework, Kane has increasingly paid more attention to them and extended the chain of inferences all the way up to a decision (Kane 1992, 2004). Bachman and Palmer (2010) and Chapelle et al. (2008) have adapted Kane's framework in somewhat different ways, but both highlight test use and consequences. In the former, assessment use and consequences are the central focus of test validation, where assessment development starts with considerations about the intended consequences. In the latter, it is seen as an inferential link from an interpretation to a decision in the validity argument, with a more elaborate discussion of the pertinent assumptions than in Kane's work.

Figure 1 illustrates the network of inferences linking test performance to a score-based interpretation and use. The first inference, *domain description,* links test tasks to test performances and is based on the assumption that test tasks are relevant to and representative of real-world tasks in the target domain. The second inference, *evaluation*, connecting test performance to an observed score, hinges on the assumptions that performance on a language test is obtained and scored appropriately to measure intended language abilities, not other irrelevant factors. The third link, *generalization,* relates an observed score to a true score and assumes that performance on language tasks is consistent across similar tasks in the universe, test forms, and occasions. The fourth link, *explanation*, connecting a true score to a theoretical score interpretation, bears on the assumption that a theoretical construct accounts for performances on test tasks. The next link, *extrapolation,* connects the theoretical score interpretation to the domain score interpretation and is based on the assumption that a theoretical construct accounts for performances on test tasks. The following link, *extrapolation,* addresses

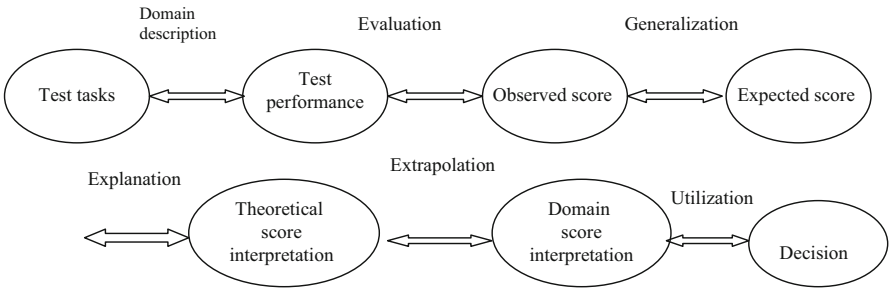


**Fig. 1** Links in an interpretative argument (Adapted based on Chapelle et al. 2008)

the connection between a theoretical score interpretation and a domain score interpretation, and the relevant assumption is that test scores reflect the quality of language performance on relevant real-world tasks. The final link, *utilization*, connects a score-based interpretation and score-based decisions. The assumptions are: test scores and other related information provided to users are *relevant, useful,* and *sufficient* for making intended decisions; the decision-making processes are appropriate; and the assessment process does not incur any negative consequences.

Corresponding to developments in the overall validation approach, validation methods have also been expanded to provide evidence in support of these inferential links. The commonly used ones for each inferential link are discussed below.

## Domain Description: Linking Test Tasks to Test Performances

Evidence supporting the domain description inference is primarily judgmental in nature, providing support that test tasks are representative samples of the domain.

Needs analysis to specify the domain and logical analysis of the task content by content specialists (Weir 1983) are typically used to establish the content relevance and representativeness of test items in relation to the domain. Corpus-based studies have recently emerged as a new technique to check the correspondence between the language used in test stimulus materials and real language use in academic settings to establish content relevance (Biber et al. 2002).

## Evaluation: Linking Test Performance to Scores

Evidence supporting the *evaluation* inference is based on the conditions under which the test is administered and the care with which the scoring rubrics are developed and applied.

### Impact of Test Conditions on Test Performance

Test conditions may impact the demonstration of intended language skills. Research has been conducted to examine the impact of test conditions on test performance to ensure that the test scores are not influenced by construct-irrelevant factors such as familiarity with computers in a computer-based test (Taylor et al. 1999). O'Loughlin (2001) examined the equivalence of scores between face-to-face and tape-mediated versions of an oral test and supported his conclusion with discourse analysis of candidates' speech elicited under the two conditions.

### Scoring Rubrics

Rater verbal protocols and analysis of a sample of test discourse (Brown et al. 2005) are commonly used to develop rubrics that are reflective of the underlying skills the test intends to elicit.

Other studies have employed quantitative methods to validate rating scales. Because a rubric with well-defined score categories facilitates consistent scoring, some studies have examined whether differences between score categories are clear using multifaceted Rasch measurement (McNamara 1996). In addition, multidimensional scaling has been applied to the development of scales for different tests and rater groups (Chalhoub-Deville 1995).

### Systematic Rater Bias Studies

Inconsistencies within and across raters are another potential source of score error in performance-based assessments. Analysis of variance and multifaceted Rasch measurement have been used to investigate the systematic effects of rater backgrounds on scores (McNamara 1996). Rater verbal protocols, questionnaires, or interviews have been employed to investigate rater orientations and decision processes (Lumley 2002). More recent studies have combined quantitative analysis of score reliability and rater self-reported data to account for rater inconsistencies (Xi and Mollaun 2006).

A related rater bias issue concerns the use of automated engines for scoring constructed response items. Automated scoring may introduce systematic errors if the scoring algorithm underrepresents the intended constructs by not including some highly relevant features or using irrelevant features. Systematic errors may also occur if the scoring model favors or disfavors certain response patterns typically associated with certain groups, and the causes for the response patterns are not related to the constructs (Carr et al. 2002).

## Generalization: Linking Observed Scores to Universe Scores

### Score Reliability Analysis

In addition to estimations of interrater reliability and internal consistency of tasks using the classical test theory, two more sophisticated methodologies have dominated score reliability studies in language testing: generalizability (G) theory and multifaceted Rasch measurement. Both methods provide overall estimates of score reliability. G theory provides useful information about the relative effects of facets, such as raters or tasks and their interactions on score dependability so as to optimize measurement designs (Bachman 2004). Multifaceted Rasch measurement is more suited to investigate the influence of individual raters, tasks, and specific combinations of raters, tasks, and persons on the overall score reliability (McNamara 1996). Given that these two techniques complement each other, studies that compared these two methods have argued for combining them to ensure score reliability (Lynch and MaNamara 1998).

In recent years, multivariate G theory has emerged as a technique to estimate the dependability of composite scores based on multiple related measures (Lee 2006) or dimensions (Xi and Mollaun 2006).

## Explanation: Linking Universe Scores to Interpretations

The *explanation* inference rests on the assumption that test tasks engage abilities and processes similar to those underlying performance on real-world language tasks indicated by a domain theory and therefore can account for performance in the domain. A wide array of methods – both quantitative and qualitative, judgmental and empirical – have been developed to gather evidence to support this assumption.

Language assessment researchers have often triangulated different methodologies to address a research topic from multiple perspectives, and this type of research has started to identify itself as mixed methods research during the last decade following its emergence as a unique research paradigm in other fields (Turner 2014).

### Correlational or Covariance Structure Analyses

Correlational or covariance structure analyses can be used to explore the empirical relationships among test items or between the test and other measures of similar or different constructs and methods for measuring them. These analyses can determine if the relationships are consistent with theoretical expectations of item homogeneity, and the convergence and discriminability of constructs and methods (Bachman 2004).

Factor analysis and structural equation modeling (SEM) are powerful techniques to test theories. Compared to experimental designs, they have the advantage of investigating a large number of variables in a single analysis.

Exploratory and confirmatory factor analytic techniques have been frequently used to confirm hypotheses or to test competing hypotheses about the factors underlying test performance, such as common abilities and processes, concurrent learning of different language skills, or common language learning interests or experiences.

SEM can not only model relationships between constructs (factors) and measured variables, subsuming confirmatory factor analysis, but also model relationships among constructs, which may represent intended abilities, other test taker characteristics, or test methods (see In'nami and Koizumi 2011 for a review).

### Experimental Studies

In experimental studies, instruction or learning interventions can be carefully planned and task features and testing conditions systematically manipulated (Bachman 1990). Therefore, they allow establishing causal effects due to treatment interventions or conditions. The effectiveness of an intervention as measured by gains in test scores attests to the soundness of the theoretical construct (Messick 1989). Research on the influence of manipulated task features on task performance can either unveil the relationship between task difficulty and task features (Iwashita et al. 2001) or disambiguate a task feature suspected to be construct-irrelevant (Xi 2005).

### Group Difference Studies

Group differences in test scores can either support score-based interpretations and uses or compromise the validity of a test for a proposed use if caused by construct under-representation or construct-irrelevant factors. Therefore, group

difference studies can test theories that groups with certain backgrounds and characteristics should differ with respect to the construct being measured. They can also forestall rival interpretations that construct under-representation or construct-irrelevant factors are associated with a test. Group differences can manifest in generalizability of scores, item or test performance (differential item, bundle or test functioning), the underlying structures of scores (differential factorial structure), strengths of relationship between the test and the criterion measures (differential criterion-related validity), or score-based decisions (differential utility) (Xi 2010).

Quantitative methods, although powerful in testing hypotheses, are limited in generating new hypotheses and do not offer insights into processes (Bachman 1990). Qualitative methods can reveal processes and strategies used by examinees during assessment, i.e., whether intended abilities and knowledge are engaged by examinees or whether any factors compromise score-based interpretations and decisions.

### Self-Report Data on Processes

Green (1997) discussed ways verbal protocols can contribute to language test validation. In cases when concurrent verbal protocols are not possible, such as with speaking tasks, stimulated recall (Gass and Mackey 2000) and retrospective interviews have been used to explore processes and strategies involved in completing language tasks.

Generally, self-report data on processes can help answer the following validity questions (Green 1997): Does the test engage the abilities it intends to assess? Do specific construct-relevant or irrelevant task characteristics influence performance? Which task types are more effective measures of the intended skills? Do different tests that are assumed to measure the same skills actually do so?

### Analysis of Test Language

Conversation and other discourse-based analyses of test language also reveal test-taking processes and strategies, although less directly than self-report data. In addition, analysis of the discourse of interaction-based tests, such as oral interviews, can inform the nature and construct of such test instruments and reveal potential construct-irrelevant factors.

Lazaraton (2002) provided a comprehensive review of studies that employ conversation analysis in language testing. Some examined the conversational features of oral interview discourse to inform understanding of the nature of the interaction as compared to that of real-life interactions. Others looked at interlocutor and candidate behavior in oral interviews and the influence of variation in interlocutor behavior on candidate performance.

Other discourse-based analytic techniques including rhetorical analysis, functional analysis, structural analysis, and linguistic analysis have been used to examine whether the distinguishing features of candidate language reflect test specifications and scoring criteria, whether oral interviews and semi-direct tests are comparable (Lazaraton 2002), or whether scores assigned by raters reflect qualitative differences revealed by discourse analysis (Cumming 1997).

### Questionnaires and Interviews

Questionnaires have been frequently used to explore test-taking processes and strategies and to elicit examinees' reactions to test tasks and the whole tests. Interviews have been used alone or in conjunction with verbal protocols of test-taking processes to follow up on interesting points (Plakans and Gebril 2012).

### Observation

Observational data on test-taking processes are usually combined with posttest interviews to reveal processes or strategies engaged by test takers or examine whether the structure of a test or process of test-taking introduces any bias. For example, O'Loughlin (2001) used observation followed up by interviews with test takers to examine the quality of the interaction between the candidate and interlocutor and identify potential bias in the way the oral interview was conducted.

### Logical Analysis of Test Tasks

This kind of analysis usually involves judgmental analysis of the skills and processes required by test tasks (Grotjahn 1986). Although experts may experience difficulty judging what an item measures, their judgmental analyses support the generation of hypotheses that can subsequently be tested by experimental or introspective studies. Logical analysis has also been used to interpret factors or to understand performance differences across groups or experimental conditions.

## Extrapolation: Linking Universe Scores to Interpretations

The *extrapolation* inference requires empirical evidence that test scores are highly correlated with scores on criterion measures. Criterion measures could be other test measures or nontest measures such as real-world language production tasks.

The relationships between tests and criterion measures are usually investigated with straightforward correlational analyses. However, selection of criterion measures that are valid indicators of performance in the domain and the reliability of them are two major issues that need to be addressed.

## Utilization: Linking Interpretations to Uses

Score-based decisions and test consequences presume and build on sound score-based interpretations. The *utilization* inference rests on several more assumptions: the score and other information provided to users are useful and sufficient, decision-making processes are appropriate, and no negative consequences are incurred as a result of the assessment process. The relevant methods are those that examine score reports and other materials communicated to users, the decision-making processes, and consequences of test use.

### Score Reporting Practices and Other Materials Provided to Users

Score reports and supplementary materials provided to score users are the only information that they base their decisions on. Therefore, care should be taken to ensure that they are useful and sufficient for decision-making. The relevant research questions are: If a composite score is reported, are the constructs measured by the components in the composite similar enough to justify aggregating the scores? Are subscores distinct and reliable enough to warrant reporting them separately? Do the subscores and/or composite scores support intended decisions? Factor analyses, generalizability studies, and subscore analyses have been conducted to address these questions (Sawaki and Sinharay 2013; Xi and Mollaun 2006). If diagnostic feedback is reported, can information about learners' strengths and weaknesses be extracted from test performance data accurately and reliably? This has been explored by applying cognitive diagnostic psychometric models to learner test item response data (Jang 2009). Moreover, the appropriateness of test performance feedback needs to be verified from stakeholders' perspectives. Interviews have been employed alone or in combination with other methods to examine stakeholders' perceptions and understanding of test performance feedback (Jang 2009).

### Decision-Making Processes

Decisions based on selection, placement, or licensure test scores usually involve setting the cut scores for minimal requirements. Although score-based interpretations may well be valid for the intended decision, inappropriate cut score models, or cut score requirements may lead to inappropriate decisions, thus compromising the utility of the test scores serving their intended purposes. Appropriate cut scores have been established based on collective judgments of a wide range of stakeholders (Sawaki and Xi 2005) or test takers' score data (Stansfield and Hewitt 2005).

### Consequences of Using the Assessment and Making Intended Decisions

Empirical research on consequences of language tests has mostly focused on washback, the impact of language tests on teaching and learning. Since the landmark Sri Lankan impact study (Wall and Alderson 1993), washback research has blossomed. Both theoretical frameworks and methodologies including interviews, surveys, classroom observations, and focus groups to investigate washback have emerged (see Alderson and Banerjee 2001, for a review).

## Problems, Difficulties, and Future Directions

In the past decade, the language testing field has advanced significantly in theoretical reformulations and expansions of the argument-based approach to validation. We have also witnessed growing empirical efforts to integrate validity evidence into a coherent argument to support a specific test use, rather than on a piecemeal basis. In the coming decade, we expect to see continuing refinements of the theoretical

approach to provide more useful guidance for and promote more rigorous conceptual thinking in practical applications.

Although the argument-based approach has increasingly been embraced for validating assessments used to make consequential decisions, endeavors to adapt it for classroom assessment and other alternative assessments with a primary focus on supporting teaching and learning have been questioned (Moss 2003; 2013). New conceptual approaches to validation emerging for these alternative assessments (e.g., Poehner 2008) could benefit from more empirical verifications to make them theoretically robust as well as practically useful.

We are in an exciting era when new conceptualizations and new technologies are pushing the boundaries of the constructs of language tests. Construct expansions have introduced new conceptual challenges and complexity in designing and validating new generations of language assessments.

## Refining the Argument-Based Approach to Test Validation

### Articulating a Clear, Coherent, and Complete Interpretive Argument

The argument-based approach offers exciting promise in guiding empirical validation research. However, applying this logical mechanism for prioritizing and organizing validation research without rigorous thinking can by no means get us as far as intended.

For each assessment use context, the interpretive argument, the network of inferences, and the pertinent assumptions must be adequately articulated through a careful logical analysis of all aspects of the assessment process. A *selective* argument driven by availability of resources and tendency to collect evidence likely to support a preferred interpretation may very likely have weak assumptions or even more seriously, weak *hidden* assumptions that are not even articulated in the argument (Kane 1992). The omission of weak assumptions in an interpretive argument in turn offers validation researchers the convenience to focus on confirming evidence in support of validity, while placing less emphasis, or ignoring potentially disconfirming evidence. This contradicts the very principles of the argument-based approach. Using an argument-based approach can by no means gloss over sloppiness in the validation efforts, or even worse, disguise attempts to cover the loopholes or weaknesses in an argument.

### Articulating Elaborated Interpretive Arguments for Typical Test Uses

The network of inferences has been fairly well developed for language tests (Bachman and Palmer 2010; Chapelle et al. 2008). However, a one-size-fits-all argument is not going to work for all assessment contexts and uses given that different uses demand different validation priorities in terms of key inferences, major assumptions associated with each inference that require backing, and types of backing needed to support each assumption. Therefore, major argument-based approaches such as Bachman and Palmer (2010) and Kane (2013) all provide generalized argument structures while emphasizing the need for adapting them to

specific test uses. However, as pointed out by Xi and Davis (2016), the level of complexity and sophistication required for constructing tailored arguments for specific uses may still discourage use among teachers and practitioners despite attempts to make it more accessible (Bachman and Palmer 2010). In addition, in the absence of a "common yardstick" for a specific assessment use (Xi and Davis 2016), it would be challenging to evaluate the completeness, coherence, and plausibility of a specific argument since the construction and evaluation of local arguments would be at the disposal of validation researchers and practitioners.

Bachman and Palmer (2010) have begun articulating assumptions in a generalized argument for assessment use; however, a particular challenge for the field is to build on these generalized assumptions and adapt them to specific assessment uses. Xi and Davis (2016) argued for developing use-specific arguments for typical test uses (e.g., admissions, licensure, placement) that specify the key validity inferences, assumptions, and types of backing needed and codifying these argument structures as part of language testing professional standards. These more nuanced, use-specific argument structures (e.g., Chapelle et al. 2008; Chapelle and Voss 2013) provide more useful guidance for practitioners to articulate and evaluate validity arguments in a specific testing context.

## Validity Research Paradigms for Different Types of Assessments

We are seeing increasing applications of the argument-based approach to the validation of language tests (Chapelle et al. 2008). However, issues have been raised regarding the fit of this approach for classroom-based assessment (Moss 2003) and for local use of standardized assessments (Moss 2013). The argument-based approach to validation may be better suited for assessment contexts where the priority is to provide useful information for score users to make decisions that have medium to heavy consequences on test takers and other stakeholders, and new validation perspectives and paradigms are needed for classroom assessments (Moss 2003) and other alternative assessments such as diagnostic assessment, dynamic assessment, and stealth assessment, for which the emphasis is on supporting learning and teaching. Diagnostic assessment focuses on the diagnosis of learner strengths and weaknesses to guide teaching and learning. Dynamic assessment attempts to make inferences about the skills that a learner possesses and his/her growth potential (Poehner 2008). Stealth assessment refers to assessment embedded in a gaming environment (Shute and Ventura 2013).

In the argument-based approach, the primary focus is on the interpretation and use of an assessment rather than the local context of use, and the problems and questions that are directly relevant to that particular context (Moss et al. 2006; Moss 2013). Validity evaluation approaches advocated by Norris (2008) and Shepherd (1993) promote a central focus on specific uses and contexts of assessment, in which validity investigations most relevant to test users are prioritized, and test users and stakeholders are closely involved in the evaluation process and use of the assessment results.

With regard to diagnostic, dynamic assessment and stealth assessment, although some work has been done to both conceptualize key validity issues and validation approaches and apply them to actual assessments (Poehner 2008, Shute and Ventura 2013), much of it is still developing and will take some time to mature and provide useful guidance for validation researchers.

## Pushing the Boundaries of Traditional Language Constructs and Validation Issues

The constructs of language tests have become increasingly more complex and may go beyond what has traditionally been defined. Such complexity in defining test constructs has been introduced by the trend towards tests of English for specific purposes, an increasing attention to English as a lingua franca (ELF) in designing second/foreign language tests, and the growing use of computers and multimedia technologies in communication. It is particularly challenging to define constructs for tests of English for specific purposes such as oral communication tests for international teaching assistants, where language skills are closely intertwined with awareness of social-cultural norms in the target classroom context, teaching skills, and content knowledge. In addition, as pointed out by McNamara (2014), the recent surge of interest in the nature and role of ELF could fundamentally change the ways in which we define communicative language ability as the target construct as well as design and validate language assessments. As technology-enhanced assessments have been on the rise, however, debates continue as to whether computer literary and digital information literacy skills (e.g., keyboarding skills) should be considered an integral component of the new construct of technology-mediated communication, feature into the construct definition for second/foreign language tests in a limited way, or as a source of construct-irrelevant variance.

These potential expansions of second/foreign language test constructs present both challenges and opportunities for us to redefine the constructs of language tests and design validation research in light of the expanded constructs.

## Cross-References

▶ Criteria for Evaluating Language Quality
▶ Critical Language Testing
▶ Qualitative Methods of Validation
▶ Washback, Impact, and Consequences Revisited

## Related Articles in the Encyclopedia of Language and Education

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

Beatriz Lado, Cristina Sanz: Methods in Multilingualism Research. In Volume: Research Methods in Language and Education

Li Wei: Research Perspectives on Bilingualism and Bilingual Education. In Volume: Research Methods in Language and Education

## References


Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (part 1) state-of-the-art review. *Language Teaching, 34*(4), 213–236.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L. F., & Eignor, D. R. (1997). Recent advances in quantitative test analysis. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Language testing and assessment, Vol. 7, pp. 227–242). Dordrecht: Kluwer Academic.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Language testing and assessment, Vol. 7, pp. 275–287). Dordrecht: Kluwer Academic.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly, 36*(1), 9–48.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on english for academic purposes speaking tasks* (TOEFL monograph series (TOEFL-MS-29)). Princeton: Educational Testing Service.

Carr, N. T., Pan, M. J., & Xi, X. (2002). *Construct refinement and automated scoring in web-based testing*. Paper presented at the 24th annual language testing research colloquium, Hong Kong, December.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33.

Chapelle, C., & Voss, E. (2013). Evolution of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 1081–1097). Boston: Wiley-Blackwell.

Chapelle, C., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

Clark, J. L. D. (Ed.). (1978). *Direct testing of speaking proficiency: Theory and practice*. Princeton: Educational Testing Service.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale: Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of Language and Education* (Language testing and assessment, Vol. 7). Dordrecht: Kluwer Academic.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621–694). Washington, DC: American Council on Education.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah: Lawrence Erlbaum Associates.

Green, A. (1997). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing, 3*(2), 159–185.

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8*(3), 250–276.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 51*(3), 401–436.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*, 135–170.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Kunnan, A. J. (Ed.). (1998). Special issue: Structural equation modeling. *Language Testing, 15*(3). London: Edward Arnold.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge: Cambridge University Press.

Lado, R. (1961). *Language testing*. New York: McGraw-Hill.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral tests*. Cambridge: Cambridge University Press.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131–166.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lynch, B. K., & MaNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–180.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31–51.

McNamara, T. (2014). 30 years on – Evolution or revolution? *Language Assessment Quarterly, 11*, 226–232.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17,* 6–12.

Moss, P. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practices, 22*(4), 13–25.

Moss, P. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement, 50*, 91–98.

Moss, P., Girard, B., & Haniford, L. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.

Norris, J. (2008). *Validity evaluation in language assessment*. Frankfurt: Peter Lang.
O'Loughlin, K. (2001). In M. Milanovic & C. Weir (Eds.), *The equivalence of direct and semi-direct speaking tests* (Series: Studies in language testing). Cambridge: Cambridge University Press.
Palmer, A. S., Groot, P. J. M., & Trosper, G. A. (Eds.). (1981). *The construct validation of tests of communicative competence*. Washington, DC: TESOL.
Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*, 18–34.
Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development*. Berlin: Springer.
Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of TOEFL iBT section scores* (TOEFL iBT research report (No. TOEFLiBT-21)). Princeton: Educational Testing Service.
Sawaki, Y., & Xi, X. (2005). *Standard setting for the next generation TOEFL*. Paper presented at the 2005 TESOL Convention, San Antonio, March.
Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.
Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Longman.
Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.
Stansfield, C. W., & Hewitt, W. E. (2005). Examining the predictive validity of a screening test for court interpreters. *Language Testing, 22*(4), 438–462.
Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–274.
Turner, C. E. (2014). Mixed methods research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1403–1417). New York: Wiley.
Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41–69.
Weir, C. J. (1983). The associated examining board's test in English for academic purposes: An exercise in content validation events. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 147–153). London: Academic.
Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22*(4), 463–508.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.
Xi, X., & Davis, L. (2016). Quality factors in language assessment. In D. Tsagari & B. Jayanti (Eds.), *Handbook of second language assessment*. Berlin: De Gruyter Mouton.
Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for TOEFL Academic Speaking Test (TAST)* (TOEFL iBT Research Report Series (TOEFLiBT-RR-01)). Princeton: Educational Testing Service.

# Qualitative Methods of Validation

Anne Lazaraton

**Abstract**

It is not surprising that there are continuing tensions between the disciplinary research paradigms in which language testers situate themselves: psychometrics, which, by definition, involves the objective measurement of psychological traits, processes, and abilities and is based on the analysis of sophisticated, quantitative data, and applied linguistics, where the study of language in use, and especially the construction of discourse, often demands a more interpretive, qualitative approach to the research process. This chapter looks at qualitative research techniques that are increasingly popular choices for designing, revising, and validating performance tests – those in which test takers write or speak, the latter of which is the primary focus of this chapter. It traces the history of qualitative research in language testing from 1990 to the present and describes some of the main findings about face-to-face speaking tests that have emerged from this scholarship. Several recent qualitative research papers on speaking tests are summarized, followed by an examination of three mixed methods studies, where both qualitative and quantitative techniques are carefully and consciously mixed in order to further elucidate findings that could not be derived from either method alone. I conclude by considering challenges facing qualitative language testing researchers, especially in terms of explicating research designs and determining appropriate evaluative criteria, and speculating on areas for future research, including studies that tap other methodological approaches such as critical language testing and ethnography and that shed light on World Englishes (WEs) and the Common European Framework of Reference (CEFR).

A. Lazaraton (✉)
Department of Writing Studies, University of Minnesota, Minneapolis, MN, USA
e-mail: lazaratn@umn.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_15

**Keywords**
Discourse analysis • Introspective techniques • Mixed methods research • Speaking/oral proficiency assessment

## Contents



## Introduction

In a recent position paper, McNamara (2011) claimed that "the distinctive character of language testing lies in its combination of two primary fields of expertise: applied linguistics and measurement" (p. 435). He further noted that language testers come to the discipline from one of two homes (rarely both): psychometrics and statistics, or applied linguistics, in which a major intellectual preoccupation is the various facets of language in use. Seen in this way, it is not surprising that there are continuing tensions between the research paradigms in which with these disciplines are situated. Psychometrics, by definition, involves the objective measurement of psychological traits, processes, and abilities and most often employs sophisticated, quantitative data collection, analysis, and interpretations. In contrast, the study of language in use, and especially the construction of discourse, often demands a more interpretive, qualitative approach to the research process. While language assessment research remains primarily a quantitative endeavor focused on product (i.e., scores), an important methodological development over the 25 years has been the appearance of qualitative research methodologies to assist in language test design, revision, and validation. This is especially so for performance testing, in which test takers speak and/or write; qualitative research techniques *have* become more prominent in the discipline in order to understand the processes and manifestations of language use in assessment, particularly tests of oral proficiency. In this chapter I trace major historical developments in qualitative language testing research, consider several research traditions that have been employed in such research, analyze a number of published studies that adopt these techniques, problematize the qualitative research endeavor in language testing, and look ahead to its future.

## Early Developments

As reported in Lazaraton (2008), from a methodological standpoint, two periods of language testing research characterize the field: pre-1990, when almost all scholarship employed a positivistic, outcome-based framework, and post-1990, after which a great deal of attention was directed to understanding the processes of performance testing. The dividing line between these periods emerged with the publication of Leo van Lier's (1989) seminal paper, in which he questioned the assumed but untested premise that the discourse produced in face-to-face, direct speaking tests involving interlocutors and test takers is, essentially, "natural conversation." He urged the language testing community to investigate not only the construct of oral proficiency but also the processes that underlie its demonstration. A second paper by applied linguists Jacoby and Ochs (1995) explored "co-construction," defined as "the joint creative of a form, interpretation, stance, action, activity, identity institution, skill, ideology, emotion, or other culturally meaningful reality" (p. 171). In response to their description of co-construction and van Lier's call for research on speaking test discourse, language testing researchers have undertaken numerous empirical studies that investigate many aspects of oral (and to a lesser extent, written) proficiency assessment.

Using qualitative discourse analytic techniques (discussed in the next section), several books published around the turn of the century investigated the nature of oral testing talk by the interviewer, the test taker, and, most notably, in the interviewer and the test candidate's "co-constructed" discourse (Lazaraton 2002). From this (and other) earlier work, we have learned that:

- Interviewers, through their talk and behavior (such as supplying answers, simplifying task directions, completing or correcting test taker responses, and rephrasing questions), bring unpredictability into the encounter, thus threatening test reliability.
- Test takers do not always produce the sorts of language or use it in ways that the test developers predict intuitively in test design.
- As a genre, language assessment interviews do "share features with conversations [but] they are still characteristically instances of interviews of a distinctive kind for the participants (Lazaraton 2002, p. 15).
- Pair and group orals, where test takers talk with one or more other test takers (instead of or addition to engaging with the interviewer), have gained popularity for several reasons: they approximate pair and group class activities; the power differential between interviewers and test takers is reduced; and a broader range of speech functions are displayed in peer talk when compared to interviewer-test taker talk.
- Pair and group test talk has been shown to be influenced by gender, personality, proficiency, and acquaintanceship, but the relationship between these variables, discourse produced, and outcome test scores is still not well understood.

In reaction to and with the help of these findings, testing organizations such as Cambridge English (www.cambridgeesol.org) instituted various refinements to their speaking tests, including the development and use of an "interlocutor frame" – an interview agenda – to guide the interviewer and the revision of rating scale descriptors to more accurately reflect the nature of test taker discourse and speech functions produced (see Taylor and Galaczi 2011 on other ways that Cambridge English has engaged in a continuous validation process for its oral assessments).

These efforts are part of an ongoing effort to keep the concept of validity and the process of validation front and center in language assessment research. According to Kane (2012), validation boils down to two questions: "What is being claimed? Are these claims warranted?" based on the evidence provided (p. 4). In a traditional sense, validity claims made by language testers are based on evidence that the assessments they design and use present a true picture of the construct being measured – for example, interactive communication or extended discourse – a task for which qualitative research techniques are ideally suited. On the other hand, an "argument-based" validation approach requires "specification of the proposed interpretations and uses of test scores and the evaluating of the plausibility of the proposed interpretative argument" (Kane, p. 3). A more detailed explication of validation is beyond the scope of this chapter; suffice it to say that "validation is simple in principle, but difficult in practice" (Kane, p. 15). The authors of the studies summarized below utilize qualitative research methods to grapple with test validation concerns for assessment interpretation and use.

## Major Contributions

In this section, I first provide background on the most widely used qualitative approach to understanding the process and outcomes of oral testing, namely, discourse analysis, followed by summaries of three recent studies that illustrate some current research in this area. Next, I overview a second qualitative research methodology that some language testers have utilized, introspective methods. Finally, I consider the principles of mixed methods research, a methodological choice that is increasingly prevalent in language assessment (LA) scholarship.

### Discourse Analysis

The most widely used qualitative approach to understanding the output of oral performance testing is discourse analysis, which traces its roots to various disciplines – primarily anthropology, linguistics, philosophy, psychology, and sociology – and can be construed broadly as an endeavor with several defining characteristics. Generally speaking, discourse analysis:

- Relies on careful transcription of authentic spoken discourse, a laborious yet fruitful part of the research process

- Accounts for and responds to the importance of context, in its broadest sense
- Produces a rich, deep analysis based on intensive engagement with the data
- Reflects one or more theories of language in use, such as accommodation theory and conversation analysis (CA)

Briefly (but see, e.g., Sidnell and Stivers 2013), conversation analysis investigates instances of "talk in interaction" about which the analyst is ideally agnostic at the outset of an investigation. The unit of analysis is the speaker turn, a central analytic construct that drives the careful transcription of the discourse to be studied. As phenomena emerge as interesting, the researcher will focus on collecting single cases that demonstrate the phenomena as well as deviant cases that should, but don't (or that do, but shouldn't). CA is insistent that assertions about participants' backgrounds, gender, and other demographic factors not be assumed as automatically relevant to the discourse being analyzed; the researcher must show how various identities are (co-) constructed, displayed, or withdrawn at a particular point in the talk. Finally, in its pure form, CA resists coding and counting data because the focus is that of a microscope looking at a single case rather than a telescope that captures phenomena at an aggregate level. Nevertheless, in subsequent sections I detail some recent research where CA data are tagged and coded in order to define and delineate constructs or to test them psychometrically. First, however, I summarize three qualitative studies that investigate facets of test discourse in face-to-face speaking tests involving one interviewer and one or more test takers.

The research of Gan (2010) and Luk (2010) centers on the production of test taker talk in group and pair orals. Gan (2010) investigated the nature of speaking test performance in higher- and lower-proficiency test taker groups in his case study of a school-based oral assessment in Hong Kong. Gan was interested in detailing the interactional features that characterized the discussions of two small groups of four secondary-level English as a foreign language (EFL) students. Gan's data consisted of fine-grained transcriptions of group discourse analyzed according to CA principles, which involved a line-by-line, sequential analysis of extended discourse. His findings revealed that the higher-proficiency group produced collaborative talk that was both "constructive and contingent": participants jointly built "opportunities for substantive conversation and genuine communication" (p. 585) with other group members in managing the conversational floor and engaging with their peer's ideas. On the other hand, the lower-proficiency group demonstrated less engagement with each other's ideas and more interactional work devoted to creating and maintaining a helpful, non-threatening discourse environment. The linguistic issues that arose in these discussions served as the basis for collaborative dialogue and assistance, but overall, their talk did not show the "contingent development of topic talk" (p. 585) that characterized the higher-proficiency group. Gan concluded that this sort of research, which focuses on the social, interactive nature of face-to-face speaking test performance, is crucial for making a validity argument about the particular assessment.

Luk's (2010) report on a group oral assessment in Hong Kong considered a different factor, that of impression management, which she defined as "a social

psychological notion that describes the process through which people consciously or unconsciously try to control the impression other people form of them so as to achieve a certain goal" (p. 27). In a very comprehensive methods of analysis section, Luk characterizes her research as "applied CA," where discourse analytic findings are supplemented with information from questionnaires and interviews. In her school-based assessment (SBA), 11 groups of four female secondary school students engaged in discussions about a text they read or a film they watched based on teacher-constructed task prompts; the classroom teacher and six students were also interviewed about their experiences. All participants also completed a questionnaire. Her findings fell into three thematic categories, including task management, content delivery, and "converging speech acts" in which conversational sequences are constructed (such as question-answer). Luk found the interactions were characterized by highly ritualized openings and closings, orderly turn taking, negotiation of meaning avoidance, and responses to fill dead air, among other features. In the end, Luk observed a "strong desire on the part of the students to maintain the impression of effective interlocutors for scoring purposes rather than for authentic communication" (p. 25). That is, participants' interactions were performative and, at times, even "collusive," where test takers rehearsed their responses beforehand. She suggests that test designers reexamine the validity of group oral assessment when test takers only speak with each other because of the possibility of planned performances.

One of the most intriguing recent papers on oral language assessment is Norton's (2013) research on speaking test talk that goes beyond previous work on the test participants to include a third, largely unexplored factor: the presence of an interlocutor frame and various testing materials. Norton employs the post-structuralist concepts of intertextuality and interdiscursivity to analyze speaking test data from two Cambridge English exams, the First Certificate of English (FCE) and the Certificate of Advanced English (CAE). Her goal was to understand the identities that speaking test participants construct by means of talking with another person and dealing with interlocutor frames and other test materials while also accounting for "the myriad of other 'voices'" present in these interactions, such as test designers who develop the assessment (p. 309). As such, her paper represents a critique of some current testing practices that were employed as solutions to problems raised in earlier investigations.

For example, past work has shown that the interviewer can be a positive, neutral, or negative factor in the discourse test takers produce; certain interviewer practices may lead to unreliable outcomes. As a result, testing agencies such as Cambridge English instituted a set of interlocutor frames that dictate what interviewers can say, so that "unscripted questions or comments should always be kept to the absolute minimum" (UCLES, 1996; as cited in Norton, p. 316). Comparing the written form of the interlocutor frame with the actual talk interviewers produced, Norton found their discourse contained "numerous deviations" from the dictated format. By including this additional, unscripted material, interviewers displayed a sort of "hybrid identity" of both teacher and examiner because "certain interviewers find it difficult to identify themselves as institutionalized, unindividuated, noninteractive subjects" (p. 316).

A further problem arose with test taker discourse: when candidates are told to say as much as they can in a set period of time, unless they are test savvy, they may confuse the need to produce a ratable sample of language with their need to be truthful. This makes it difficult to distinguish "cannot talk" vs. "will not talk" candidates who respond truthfully rather performatively by engaging verbal behaviors that may be appropriate in conversation but are detrimental to ratings of test talk. Norton's evidence strongly suggests that task design itself is implicated in co-construction of performance and must be accounted for as such; she concludes that it is "intrinsically problematic . . . to impose such a framework to elicit language for assessment purposes when the framework itself may limit participation in speaking tests in ways which cannot be easily predicted" (p. 325). She recommends that testing organizations ensure that all candidates understand assessment criteria, including the desirability of initiating topics and expanding on answers to produce a sufficient sample of language for rating purposes.

## Introspective Techniques

While these discourse analytic studies looked at the language produced in the speaking test context, another qualitative research has explored the cognitive processes in which raters engage when assessing language production in performance tests. Often this work utilizes *introspective methods*, which aim to generate usable data on cognition during or after a particular task. Sasaki (2014; see also Green 1998) describes such techniques, including *think-alouds*, where participants articulate their thoughts while engaged in a task, and after-the-fact *recalls* which can be stimulated with a memory aid or collected alone. Sasaki contends that analyses of these verbal protocols are ideal for complementing, rather than merely supplementing, more quantitative analytic techniques. In other words, such inquiry "can also contribute to knowledge accumulation in the LA [language assessment] field by adding a harvest of studies with nonpositivist perspectives that are quite different from those that have hitherto prevailed in the field" (p. 16).

Three such studies are illustrative. In research looking at how oral proficiency ratings may be influenced by rater accent familiarity, defined as "gained through having learned the first language (L1) of the test takers as an L2 in the past" (p. 770), Winke and Gass (2013) asked 26 trained raters to assess Internet-based Test of English as a Foreign Language (TOEFL) speaking test samples and then reflect on their rating processes. The raters, who were native speakers (NS) of Chinese, Korean, or Spanish, engaged in 20–30 min stimulated recalls while viewing their rating sessions. The authors analyzed the recalls using an analytic inductive approach, which generated a total of eight themes, three of which were related accent familiarity: the test takers' L1, the test takers' accent, and the raters' heritage status. They concluded that "although sensitivity to test-taker accents seemed to occur naturally in the rating process, findings suggest that when raters have learned or know, to varying degrees, the test takers' L1, they tend to orient themselves to the speech in a biased way, compromising test reliability" (p. 762).

Rater thought processes were also analyzed by May (2011), who employed stimulated verbal recalls to understand the interactional features that were salient to raters of 12 paired speaking tests. Along with rater notes and data from rater discussions, her analysis indicated that a number of features were noticed, including interpreting and responding to another's message, working together, and adding to the authentic interaction taking place. May's concern was that it is difficult to evaluate an individual's performance in a co-constructed interaction; as a result, paired orals are not necessarily a panacea to the problems encountered in more traditional interviewer-test taker assessments. In other words, pair and group oral assessments have potential validity problems of their own.

Finally, rater cognition is also of interest in writing assessment. Li and He (2015) utilized primarily qualitative, introspective methods in their study of an analytic and a holistic scale used by nine raters of ten essays produced for the Chinese College English Test. The authors focused on how the rating scale type appeared to influence rating strategies and the textual features focused on by raters. Li and He used think-aloud protocols, questionnaires, and semi-structured interviews to collect their data. They found that holistic scales led to more interpretation strategies; judgment strategies were more prevalent with the analytic scale. Overall, the authors suggest that holistic scales force raters to focus on more limited set of text features and to adopt essay comparisons strategies. Additionally, the lack of detailed descriptors in the holistic scale led to rater difficulties in defining and assessing the construct, leading to less reliable and valid scoring.

## Work in Progress

A more recent scholarly trend, increasingly apparent in the last 10 years, is the emergence of mixed methods research (MMR) in language assessment. MMR involves the conscious, principled mixing of quantitative and qualitative methodologies and analyses, and in language assessment, to arrive at a more comprehensive understanding of performance test factors in test talk, rater cognition, and the perspectives and beliefs of test stakeholders. The roots of MMR can be traced to the concept of "triangulation" in qualitative research (see Creswell 2014, an essential source on MMR); as a research strategy, triangulation involves one or more of the following: obtaining multiple sources of data, including multiple groups of participants, and employing multiple research techniques. According to Turner (2014), a basic premise of MMR is that qualitative and quantitative researches are not incompatible, but complementary in their strengths and weaknesses. Although mixing methods has been going on for a long time, it wasn't until around 2003 when a set of guiding principles started to take shape. Nevertheless, despite "increasing evidence of [LT] research employing both qualitative and quantitative approaches, [but] specific articulation of employing an MMR is still rare" (Turner 2014, p. 4). Both Turner and Brown (2014a) articulate various design types for MMR, each of which is a permutation of temporal elements – concurrent and sequential – in terms of data collection and analysis and of research goals as exploratory or explanatory.

However, different authors array these factors using sometimes dissimilar vocabulary, so it is not always a simple manner to compare research designs in the absence of informative visuals, as the two studies described below include.

One of the first published MMR research papers in language testing is Kim's (2009) examination of the differences in proficiency judgments between native speaker and nonnative speaker (NNS) teachers of English. The impetus for his research was that previous quantitative analyses of rater behavior and English language background were not sufficiently fine-grained. A total 12 Canadian and 12 Korean teachers of English rated semi-direct speech samples (where test takers speak into a recorder rather than with a person) from ten college-level English as a second language (ESL) students performing a total of eight speaking tasks. Rater behavior was analyzed using multifaceted Rasch; rater comments on student performance were analyzed qualitatively. Teacher comments were open coded, resulting in 19 recurring criteria. Kim's findings indicated that both groups of teachers showed good internal consistency in their ratings and a similar harshness pattern. It was in the comments about evaluative criteria where the rater groups exhibited notable differences. The Canadian teachers produced a larger number of comments, and they were more detailed and elaborate than those from the Korean teachers, although both groups were most concerned with vocabulary, pronunciation, and overall language use. Kim cautiously interprets his findings regarding the qualitative differences he detected, hypothesizing that nonnative speakers (NNS) aren't always trained to assess details of performance and may come from different evaluation cultures. In any case, what is most notable is the degree to which Kim carefully explains his research design, sampling, and analytic techniques using accessible terminology and provides a very helpful diagram of the research procedures.

Youn (2015) is a second exemplary inquiry employing mixed methods to develop a validity argument for assessing second language (L2) pragmatics. His research used discourse data from open role-plays to inform task design and rating criteria development and also analyzed rater performance with FACETS. A total of 102 students and four native speakers engaged in role-plays, discourse from which was transcribed and analyzed using CA methods in order to "back[ing the] valid task design and sound rating criteria assumptions" (p. 203). Five "interaction-sensitive, data-driven" rating criteria were derived from this analysis: contents delivery, language use, sensitivity to situation, engaging with interaction, and turn organization, each of which Youn illustrates with a relevant data fragment that depicts how the rating criteria were derived. In Youn's words, "the CA findings helped examine a degree of authenticity and standardization of the elicited performances along with detailed descriptions for rating criteria" (p. 203); "the mixed methods generated convincing backing for the underlying assumptions of the evaluation inferences" (p. 218). Like Kim's work described above, Youn presents two informative diagrams and figures: the first represents the evaluation inference schematically, and the second depicts the study design.

Finally, mixed methods were also used by Zhao (2013) to develop and validate rubric for measuring authorial voice in L2 writing. Four raters assessed authorial voice in 200 TOEFL iBT writing samples using a preliminary rubric containing

11 features rated on a 0–4-point scale. In the development phase, the author used principal components analysis of the resulting ratings to generate a set of construct dimensions to inform the creation of the scoring rubric. She also collected and analyzed think-aloud protocols and interview data "to supplement the quantitative analysis and provide additional evidence on rubric reliability, applicability, and construct validity" (p. 205). The qualitative data were then used to create the final authorial voice rubric used in the validation phase of the study. Her findings indicated that both the quantitative and qualitative data supported a three-dimensional conceptualization of voice. Zhao concludes that a rater's thoughts and feelings about the overall quality of voice in a writing sample are less a matter of the *quantity* of individual voice elements in the text; more meaningful to raters is *how* they are used.

As promising as mixed methods research appears to be for language assessment, there are challenges associated with the approach that testers must consider.

## Problems and Difficulties

There are indeed unique concerns in MMR research that must be accounted for. For one, it is not always clear how research questions should be formulated and ordered or prioritized: Separately? Sequentially? Overarching? And what is sampling process? How should MMR be evaluated? And who has the expertise to engage in both qualitative and quantitative researches (Turner 2014, pp. 10–11)? In any case, as language assessment research designs have become more complex, it is even more important for scholars to include visuals that represent a sometimes nonlinear research process; this requirement should be taken more seriously in presenting and publishing MMR research.

Along with the inclusion of schematics and visuals, it is essential for researchers to explicate clearly the framework being employed. Triangulation as a research strategy is a very good one, but simply including multiple methods in an investigation is not the same as engaging in rigorous mixed methods research. "MMR uses a specific logic, especially the *fundamental principle of MMR*, i.e., 'the research should strategically combine qualitative and quantitative methods, approaches, and concepts in a way that produces complementary, strengths and nonoverlapping weaknesses'" (Johnson et al. 2007; as cited in Brown 2014a, p. 9; emphasis in original). Brown astutely notes that "if the qualitative methods and quantitative methods are simply used simultaneously or sequentially, with them not interacting in any particular ways, the research might be more aptly labeled multimethod research" (p. 9).

More broadly, evaluative criteria for qualitative and mixed methods language testing research must be developed and/or refined. As a research community, language testers have a long history of evaluating positivist, quantitative research according to established criteria such as validity (the current intellectual preoccupation; see Kane 2012), reliability, replicability, and generalizability. We also have some familiarity with their qualitative counterparts of dependability, credibility,

confirmability, and transferability, but there is still misunderstanding of and debate about how to weigh and describe these criteria – or if they are even the right criteria. Enter mixed methods, an even more complicated endeavor. Brown (2014a) maintains that evaluative criteria for both quantitative and qualitative researches have thematic parallels: *consistency* captures both reliability and dependability; validity and credibility are both concerned with *fidelity*; *meaningfulness* characterizes generalizability and transferability; and *verifiability* subsumes both replicability and confirmability (p. 119). While this heuristic is certainly helpful in setting out the correspondences between quantitative and qualitative research, it is uncertain how, and to what degree, language testers understand these meta-concepts, much less the central and complex issue of validity in mixed research, "legitimation," which "is to MMR what *validity* is to quantitative research and *credibility* is to qualitative research" (Brown 2014a, pp. 127–128).

## Future Directions

One area that continues to be fertile ground for language assessment research is in understanding the consequential validity of language tests. Surprisingly, impact studies of "their uses, effectiveness, and consequences" (Shohamy 2001, p. xvi) are not plentiful (but see, e.g., some of the papers in Shohamy and McNamara 2009; and O'Loughlin 2011, on the use and interpretation of IELTS scores in university admissions). Should critical language testing (CLT) be considered one type of qualitative research? It seems clear that a positivist paradigm, where objectivity and generalizability are valued goals, is not really consistent with the subjective, "lived" experiences that CLT would tap. Shohamy (2001) claims that numbers are symbols of "objectivity, rationalism, … control, legitimacy and truth" (p. x) and their power lies in the fact that they can be challenged only by using different numbers to counteract them; testers "own" the numbers. From this perspective, CLT and quantitative inquiry may well be incommensurable, but this position is unsupported by evidence and is only personal conjecture at this point.

Other research approaches, such as ethnography, are ideal for shedding light on classroom-based assessment (CBA) practices. For example, Hill and McNamara's (2012) ethnography of one primary and one secondary Indonesian as a foreign language classroom in Australia focused on assessment processes, especially in terms of the evidential, interpretive, and use dimensions and scope of CBA. The authors collected and analyzed "a diverse range of data" that established the processes in which classroom teachers engage, the materials from which they gained assessment information, and their views on language learning and assessment. Their data also shed some light on assessment from the learners' perspectives. A different sort of classroom was the locus for an ethnographic study by Tsagari (2012), who examined First Certificate in English (FCE) test preparation courses in Cyprus with the aim of explicating the "details of teachers' instructional behaviors and … descriptions of classroom practices" (p. 37) in order to understand the potential washback of the courses. Fifteen classroom lessons totaling 24 h of observation data

across three schools along with other supplementary information were analyzed and then presented as data fragments. Tsagari's findings point to both positive and negative washbacks in the FCE test preparation classes. The amount of work dedicated to reading, listening, speaking, and writing was seen as positive impact, while the reading test format, the limited genre writing, and test-wise listening strategies narrowed the authenticity and applicability of the strategic practice in which students engaged.

Additionally, although there has been frequent talk about the role of World Englishes (WEs) in language assessment (see Brown 2014b), to date there is not much empirical inquiry that takes up WEs in a systematic way. Harding (2014) suggests that a "guiding principle of new research on the communicative competence construct must be a focus on "adaptability" . . .to deal with different varieties of English, appropriate pragmatics, and fluid communication practices of digital environments" (p. 194). In the global English context, Harding argues that the research agenda must include "the development and validation of language tests that specifically assess a test-taker's ability to deal with diverse, and potentially unfamiliar varieties of English. These tests would use as their basis a different range of skills and abilities including: ability to tolerate different varieties of English" (p. 194). He notes that paired and/or group speaking assessments, which may require lingua franca interaction, could provide such evidence; "discourse data yielded from tasks of this kind (complemented by stimulated recall performed by test-takers) could be analyzed with a view to locating points at which these abilities are tapped in these interactions" (p. 195).

One final area where qualitative research is underrepresented relates to the Common European Framework of Reference (CEFR). McNamara (2014) points out that while there are "a plethora of studies on applications of the CEFR in various contexts. . . few of these studies are critical in any important sense. Most are overwhelmingly and unquestionably technist and functionalist" (p. 228).

McNamara further criticizes much CEFR research because it fails to even mention English as a Lingua Franca (ELF); more broadly, "the lack engagement with the larger question of the role and function of the CEFR" (p. 229) is indicative of the lack of engagement with larger sociopolitical issues that are endemic in language testing. We can hope that such inquiry relies, at least in part, on qualitative research techniques that hold much promise for delving more deeply into test impact and consequences. If qualitative research techniques can be properly combined with more traditional quantitative methodologies that lend themselves capturing the scope of language assessment phenomena, all the better.

## Cross-References

▶ Critical Language Testing
▶ Criteria for Evaluating Language Quality
▶ Methods of Test Validation

## Related Articles in the Encyclopedia of Language and Education

Beatriz Lado; Cristina Sanz: Methods in Multilingualism Research. In Volume: Research Methods in Language and Education

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

Li Wei: Research Perspectives on Bilingualism and Bilingual Education. In Volume: Research Methods in Language and Education

## References

Brown, J. D. (2014a). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.

Brown, J. D. (2014b). The future of World Englishes in language testing. *Language Assessment Quarterly, 11*(1), 5–26. doi:10.1080/15434303.2013.869817.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Los Angeles: Sage.

Gan, Z. (2010). Interaction in group assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585–602. doi:10.1177/0265532210364049.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly, 11*(2), 186–197. doi:10.1080/15434303.2014.895829.

Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*(3), 395–420. doi:10.1177/0265532211428317.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171–183. doi:10.1207/s15327973rlsi2803_1.

Kane, M. (2012). Validating score interpretations and uses: Messick Lecture. *Language Testing, 29*(1), 3–17. doi:10.1177/0265532211417210.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217. doi:10.1177/0265532208101010.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Language testing and assessment 2nd ed., Vol. 7, pp. 197–209). New York: Springer.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly, 12*, 178–212. doi:10.1080/15434303.2015.1011738.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly, 7*(1), 25–53. doi:10.1080/15434300903473997.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127–145. doi:10.1080/15434303.2011.565845.

McNamara, T. (2011). Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435–440. doi:10.1177/0265532211413446.

McNamara, T. (2014). 30 years on – Evolution or revolution? *Language Assessment Quarterly, 11*(2), 226–232. doi:10.1080/15434303.2014.895830.

Norton, J. (2013). Performing identities in speaking tests: Co-construction revisited. *Language Assessment Quarterly, 10*(3), 309–330. doi:10.1080/15434303.2013.769549.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146–160. doi:10.1080/15434303.2011.564698.

Sasaki, M. (2014). Introspective methods. In A. Kunnan (Ed.), *Companion to language assessment.* Wiley. doi:10.1002/9781118411360.wbcla076.

Shohamy, E. (2001). *The power of tests: A critical perspective on the use of language tests*. New York: Pearson.

Shohamy, E., & McNamara, T. (2009). Language assessment for immigration, citizenship, and asylum. *Language Assessment Quarterly,6*(4). doi:10.1080/15434300802606440.

Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. West Sussex: Wiley-Blackwell.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 171–233). Cambridge: Cambridge University Press.

Tsagari, D. (2012). FCE exam preparation discourses: Insights from an ethnographic study. *UCLES Research Notes, 47*, 36–48.

Turner, C. (2014). Mixed methods research. In A. Kunnan (Ed.), *Companion to language assessment*. Wiley. doi:10.1002/9781118411360.wbcla142.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508. doi:10.2307/3586922.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762–789. doi:10.1002/tesq.73.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199–225. doi:10.1177/0265532214557113.

Zhao, C. G. (2013). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing, 30*, 201–230. doi:10.1177/0265532212456965.

# Training in Language Assessment

Margaret E. Malone

**Abstract**

This chapter examines the major issues involved in providing appropriate training and professional development for language instructors to improve their knowledge and skills and make informed decisions throughout all aspects of the assessment process. First, the chapter reviews the major approaches to conducting language assessment within the context of educational policies and societal beliefs over time. In reflecting on the changing contexts and approaches, the chapter identifies the underlying philosophies of training in assessment and how such philosophies align with approaches to both assessment practice and how teachers have learned how to conduct assessment. The chapter also investigates different approaches to training. Traditional language testing textbooks and their content, as well as the recent increase in the availability of such textbooks, are highlighted. The chapter then turns from traditional textbooks to traditional professional development and addresses ways that such professional development has changed, including the availability of distance learning and other online resources, as well as how such approaches provide opportunities for innovation and improved understanding of language assessment. Finally, the chapter addresses the ongoing challenges in training for teacher professional development, including lack of resources and frequent lack of agreement between language testers and language teachers regarding the most essential topics for teachers to understand in learning about, developing, selecting, and using language assessments with their students.

M.E. Malone (✉)
Assessment and Evaluation Language Resource Center, Department of Linguistics, Georgetown University, Washington, DC, USA
e-mail: malonem@georgetown.edu; mmalone@cal.org; meg@cal.org

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_16

**Keywords**
Language assessment literacy • Teacher professional development • Assessment

## Contents



## Introduction

This chapter updates and addresses some of the major issues in training language instructors to make informed decisions in all aspects of the assessment process; in this context, the "assessment process" refers to developing, scoring, interpreting, and improving classroom-based assessments developed by language instructors as well as selecting, interpreting, and sharing results of large-scale tests developed by professional testing organizations (Stoynoff and Chapelle 2005; Bachman and Palmer 1996). Within the context of providing training in language assessment, this chapter explores "language assessment literacy" (Taylor 2013; Inbar-Louie et al. 2013; Malone 2013; Stiggins 1997; Stoynoff and Chapelle 2005; Boyles 2005), discusses expanded definitions of assessment literacy, and reviews the available resources for training in language assessment, as well as work that still needs to be done.

As pressure for language instructors and educational institutions to provide information on students' progress has increased since the 1880s and skyrocketed in the past decade (Llosa 2011; Brindley 1997), attention has focused on the testing that takes place within the context of language teaching and learning. The 2001 passage of *No Child Left Behind* (NCLB) in the United States mandates annual assessment of the English language proficiency of all English language learners enrolled in elementary and secondary programs and emphasizes the need to track and monitor student outcomes and progress in both English language and content areas (Alicea 2005). Although Europe and other countries do not mandate the use of the Common European Framework of Reference for Languages (CEFR), in that member nations are not required to adopt it or its aligned tests, by emphasizing language teaching and learning (Little 2012), the CEFR exerts great influence on the teaching and assessment of language (Davies et al. 1999) in Europe and beyond, thus demonstrating one way that language assessment has increased in importance in many places in the world.

Despite the growth of standards-based education, standards for teacher certification, and an increase in tests administered, there is no clear framework of what is required or even needed for language instructors to reliably and validly develop, select, use, and interpret tests or the extent to which these standards are used for classroom assessment (Llosa 2011). Therefore, the issue is how to identify the best approaches for support and training for those who "have to do the real work of language teaching" (Carroll 1991, p. 26) when they assess their students.

In addition to the practical and pedagogical concerns about teacher assessment knowledge and skills, the political arena also influences how, when, and why students are assessed. With the arrival of NCLB in the United States and the CEFR in Europe and beyond, assessment of language learners' progress has only strengthened in political, practical, and pedagogical importance. This chapter examines how the underlying philosophies of training in assessment have changed over time, in response to societal and educational changes in policy and practice. It also examines how different approaches for training in language assessment, from textbooks to distance learning, have altered such training. Finally, it examines ongoing challenges and future directions for increasing the "assessment literacy" of language instructors for the improvement of language learning and teaching.

## Early Contributors

Like education, language assessment is a microcosm of what is happening in larger society. This part of the chapter describes the three early periods of language testing (1800s–1980s) and discusses how each period's philosophies were reflected in available assessment training. Spolsky (1977) has divided language testing from the 1800s through the 1980s into three major periods: prescientific, psychometric, and sociolinguistic.[1] The prescientific approach, as practiced in the United States and Europe, relied mainly on the judgments of instructors as they assessed a translation, composition, or oral performance or another open-ended task presented to students. The very term "prescientific" judges this approach "unscientific"; the lack of science as applied to language testing during this period resulted in debates as to the reliability of written and oral exams administered to large groups of students and rated by different instructors. The literature does not reveal any systematized, required training for instructors on how to develop the questions for these tests, guidelines for rating the test results, or available training for the instructors in rating the examination performances.[2] As far back as 1888, debates ensued as to the reliability of these written (or oral) exams, administered to large groups of students and rated by different instructors with varying understanding of expected outcomes

[1]Spolsky (1981), Barnwell (1996), and others have alternative names for these periods; this paper uses the original terms.

[2]While some large-scale tests for admittance to universities or professions included oversight by committees, there is no evidence of such oversight for classroom assessment.

(Spolsky 1995). Despite these criticisms, it is important to note that such exams, including professional exams for admittance to, for example, the Indian Civil Service Exam supplemented patronage for candidates to the civil service. In other words, early language tests, though their developers and raters may have lacked rigorous formal training in language assessment, were often viewed as a more democratic means of admitting students to university and the workplace than simply using personal connections (Spolsky 1995).

By contrast, the second period, termed as the psychometric period, emphasized statistics and measurement and moved away from open-ended test questions to test items focusing on discrete aspects of language, such as vocabulary, grammar, pronunciation, and spelling. The format for testing also changed from the first to the second period, while in the prescientific period, students may have responded to prompts for a written essay or oral response and test items in the psychometric period included more, but shorter, questions. It was at this time that item types such as multiple choice, true/false, and similar short questions gained popularity in testing. The popularity of this approach was thus reflected in course offerings at institutions of higher education; Jonic (1968), as cited by Spolsky (1995), reports that, by 1920, courses in educational measurement were being offered by most US state universities, although such educational measurement approaches had not yet spread to language learning.

Therefore, the shift from fewer test items with long responses that took time to score to more test items with short, easy to score test items, was underway. While this new phase in language testing addressed some of the criticisms of the prescientific phase, it introduced new challenges. Despite Jonic's (1968) reference to the development and availability of educational measurement courses, there is no indication that such courses were uniformly required of teachers; therefore, the change was not accompanied by a similar change in approach to language testing courses. During this period, the work of testing and teaching was divided; testing organizations developed large-scale tests to measure student progress, and teachers provided instruction to students (Stoynoff and Chapelle 2005). Therefore, a gulf developed between instructors and test developers.

By the 1970s, changes in society, educational measurement, and theories of language learning resulted in a shift toward the sociolinguistic period.[3] During this period, the focus shifted from discrete-point testing toward tests to measure meaningful communication (Ommagio 1986). A great deal of literature is devoted to how language instructors should (and should not) be trained to assess according to variations of this approach (Bachman and Savignon 1986; Lantolf and Frawley 1985). One of the most popular approaches to assessing communicative competence during this period in the United States was the *ACTFL Proficiency Guidelines*, while later in Europe, work began on what would become the *Common European Framework of Reference*. By the early 1980s, training in various approaches to assessing

[3]Canale and Swain (1982) and others refer to this as "communicative competence" or "the proficiency approach" (Barnwell 1996).

communicative competence became available, and language instructors could seek and receive training in various approaches. As this period in testing spread into the 1980s, educational reform in the United States and efforts by the Council of Europe to reform language teaching prodded the sociolinguistic movement toward measuring outcomes based on shared standards for language learning (Stoynoff and Chapelle 2005).

However, the gap in skills held by teachers and test developers that developed during the psychometric period tightened during the sociolinguistic period and narrowed further with the introduction and incorporation of standards in the language classroom. With the 1980s and 1990s, a new era of language testing, with roots in the education reform movements in Europe and the United States, emerged.

### Current Trends

Spolsky (1995) and others have described thoroughly the three early periods in modern language testing. Following and overlapping the sociolinguistic period, the literature shows an increased emphasis on authentic, performance (or outcomes-based) assessment to reflect what students need to do with the language in real-life settings (Wiggins 1994) as well as an increased importance on shared, common standards with which to assess students. During this time, methods of collecting information from students gained popularity, such as portfolios of student work and student self-assessment, and increased emphasis on the authenticity of the task the student was to perform with respect to language use in daily life (Moore 1994). In the 2000s, emphasis on testing, including language testing, has steadily increased. The release of the CEFR in Europe and beyond and the passage of NCLB, as well as the introduction of the Common Core State Standards Initiative in the United States, have only magnified the importance of testing worldwide. The connection between assessment, standards, and politics highlights the importance of training language instructors so that they can adequately assess their students' progress toward local, national, and/or international goals and standards.

## Major Contributors

Any history of language testing will readily name a number of influences on language assessment; it is more difficult to pinpoint at what point changes in the language testing arena begin to influence the pre- and in-service training of classroom teachers because of the gradual nature of the change. The impetus for the three periods described in the previous section began with primarily large-scale assessments, such as admission to university and professions; the rate at which results and lessons learned from large-scale assessments trickle down to instructors and into preservice teacher texts is unclear and undocumented. This emphasis is reflected not only in the volume of assessments available throughout the world but also in the number of texts available for training instructors in assessment. Reviewing the three

periods is important to contextualize how training for language assessment has evolved over the past two centuries. During the prescientific period, the assessment role fell largely on individual instructors, while during the psychometric period, test development was largely in the hands of expert psychometricians, and thus language teachers did not receive much, if any, training in language test development. However, the sociolinguistic period represented a time when language teachers began to become increasingly involved in language testing. The impact of the sociolinguistic period is evidenced by the titles and content of texts developed on language testing over a 40-year period. In this section, I will address two major contributions to training in language assessment: traditional text-based materials and technology-mediated materials and information that became available in the 1990s and beyond.

## Text-Based Materials

There are several ways to examine language testing textbooks, including length, content, and quantity of available textbooks. Cohen (1994) references seven other textbooks on language testing available at the time of printing and points out that there were not as many available in the edition published 15 years earlier. This gap shows the crux of the issue of training in language assessment; during the psychometric period, "large-scale standardized instruments [were] prepared by professional testing services to assist institutions in the selection, placement and evaluation of students" (Harris 1969, p. 1), and the focus was on training professionals to develop items for standardized tests rather than training language instructors to assess their students. Examining the bibliographies of over 560 language testing texts, the author initially selected ten published from 1967 to 2005 to contrast on page lengths and number of citations listed in Google Scholar and then three more published or revised from 2005 onward. Table 1 shows these results.

While this table includes only a very small sample of textbooks available in language testing from the late 1960s until present, it shows differences and similarities over time. For example, while Valette and Harris were contemporaries, the lengths of their textbooks were different, and Valette had nearly four times as many references as Harris. In 2005, Harris has twice as many citations on Google Scholar as Valette; 10 years later, his Google Scholar citations dwarf hers. In addition to the contrasts between specific texts, there are definite changes over time. First, text length increased over time, as knowledge about language testing grew, and, similarly, the number of references included in texts increased. It is also interesting to note the contrast between the number of Google Scholar citations for each text in 2005 and 10 years later is remarkable. This growth first speaks to the increased power of the Internet in general and Google Scholar in particular of tracking citations and secondly shows how much more frequently all sources are cited even 10 years later. Table 1 also shows how the numbers of pages and the number of references have increased over time

**Table 1** Distinctions in page lengths and number of references in language testing books

| Author/text | Date of publication | Page length | Number of references | Citations on Google Scholar (2005) | Citations on Google Scholar (2015) |
|---|---|---|---|---|---|
| Harris, D. *Testing English as a Second Language* | 1969 | 146 | 7 | 40 | 679 |
| Valette, R. *Directions in Foreign Language testing* | 1967 | 200 | 26 | 18 | 21 |
| Oller, J. W. *Language Tests at School* | 1979 | 421 | 370 | 140 | 1,209 |
| Cohen, A.D. *Testing Language Ability in the Classroom* | 1980 | 132 | 172 | 56 | 153 |
| Henning, G. *A Guide to Language Testing* | 1987 | 158 | 117 | 37 | 506 |
| Hughes *Testing for Language teachers* | 2003 | 154 | 66 | 343 | 3,030 |
| Bachman, L. *Fundamental Considerations in Language testing* | 1990 | 359 | 751 | 751 | 6,477 |
| Weir, C. *Understanding and Developing language tests* | 1995 | 170 | 83 | 65 | 583 |
| Brown, H.D. *Language Assessment*: *Principles and Classroom Practice* | 2004 | 160 | 302 | 9 | 1,511 |
| Stoynoff and Chapelle *ESOL Tests and Testing*: *A Resource for Teachers and Program Administrators* | 2005 | 204 | | 1 | 45 |
| Bachman and Palmer. *Language Assessment in Practice*: *Developing Language Assessments and Justifying Their Use in the Real World* | 2010 | 493 | 193 | n/a | 281 |
| Fulcher, G. *Practical Language Testing* | 2010 | 346 | 377 | n/a | 97 |
| Carr, N. *Designing and Analyzing Language Tests* | 2011 | 361 (plus CD appendix) | | n/a | 31 |

In preparing this chapter, the author examined over 100 language testing publications, including books, articles in peer-reviewed journals, and guidelines. In addition to the gap between page length and number of citations that exists between various texts, there is also a difference between earlier and later editions of texts, as

**Table 2** Changes in Hughes', Cohen's, and Bachman and Palmer's textbooks

| Author and book title | Date of publication | Page length | Number of references |
|---|---|---|---|
| Hughes<br>*Testing for Language Teachers* | 1989 | 154 | 66 |
| Hughes<br>*Testing for Language Teachers* | 2004 | 217 | 186 |
| Cohen<br>*Testing Language Ability in the Classroom* | 1980 | 132 | 172 |
| Cohen<br>*Testing Language Ability in the Classroom* | 1994 | 362 | 433 |
| Bachman and Palmer<br>*Language Testing in Practice*: *Designing and Developing Useful Language Tests* | 1996 | 370 | 88 |
| Bachman and Palmer<br>*Language Assessment in Practice*: *Developing Language Assessments and Justifying Their Use in the Real World* | 2010 | 493 | 193 |

Cohen points out. Therefore, Table 2 shows the differences in Hughes', Cohen's, and Bachman and Palmer's textbooks over time.

The differences in length and references mirror additions of content to the text. While all texts referenced above include steady reminders of reliability, validity, and practicality, the 1990 and onward versions include more references to assessments such as portfolios and other practices that became widespread in the 1980s. In addition, Hughes added a chapter on assessing children because of the increased emphasis on testing this age group (Hughes 2004). Cohen (1994) and Bachman and Palmer (2010) more than doubled the number of references, suggesting that teachers required more information in the 15 years that passed between publications. Bachman and Palmer also adapted their title from *Language Testing in Practice*: *Designing and Developing Useful Language Tests* to *Language Testing in Practice*: *Developing Language Assessments and Justifying their Use in the Real World*. The shift in title shows the emphasis on assessment rather than testing and the growing emphasis of "real-world" use of assessment. As assessments change, the textbooks used in teacher training must change as well.

Just as new language testing textbooks began focusing on classroom teachers' practical needs, additional text-based resources emerged in the 1990s and have continued to be used in the field. While early textbooks often combined theoretical explanations with samples from actual assessment practices, the 1990s saw an explosion of textbooks that could supplement existing ones by supplying examples that could be readily included in the classroom or a "how to" on classroom assessment.

O'Malley and Valdez Pierce's (1996) *Authentic Assessment for English Language Learners*: *Practical Approaches for Teachers* represented a new approach to language testing textbooks; it combines theory and practice in an accessible volume for classroom teachers. Its rubrics, checklists, and practical advice on applications can easily be incorporated into the classroom. At a similar time, Brown (1998) produced

a volume with 18 different activities with input from three to eight international contributors for each activity type. Others in language testing also worked to model and explain solid theories of language testing coupled with practice; Bachman and Palmer (1996) and Genesee and Upshur (1996) published textbooks on language testing with an emphasis in both their titles and tone toward classroom teacher use. Unlike traditional language testing textbooks, both volumes emphasized the specific issues and problems faced by classroom teachers and aimed to combine a theoretically strong approach to language testing with practical help. For example, Genesee and Upshur (1996) include conferencing and portfolios, both approaches that gained popularity in the 1990s, as well as tables that describe the benefits of portfolios.

In the spirit of combining the information of a language testing textbook and the practicality of a "how-to" manual for teachers, Davidson and Lynch (2002) have produced *Testcraft*: *A Teacher's Guide to Writing and Using Language Tests*. Their approach emphasizes the importance of developing solid test specifications based on language testing research. At the same time, they tackle practical issues of teamwork in the test development process and ways to approach inevitable conflicts, as well as including scenarios applicable to situations their readers may encounter. Few language testing texts address the importance of teamwork and the challenges inherent in working with colleagues who have differing viewpoints about the purposes and uses of the test as well as suggest approaches for addressing not just the content of such issues but also working with colleagues.

Stoynoff and Chapelle (2005) published *ESOL Tests and Testing*, a volume which includes reviews of common English language tests, as well as chapters on the "basics" that language instructors should know before using any test. Stoynoff and Chapelle stress the importance of making informed decisions in all aspects of the testing process, and the structure of the volume supports this approach. The reviews are embedded in the book, rather than appearing at the beginning or the end, and this sequence emphasizes the importance of contextualization in test selection. This volume points to the issue of "assessment literacy" in language instructors and the need to provide practical and usable resources to language instructors to ensure that tests are selected and used properly.

Bachman and Palmer (2010) updated their original 1990 book, and it is widely used. In addition, the slight change to the title emphasizes the use of testing in "real-world" situations and the decisions made on the basis of language tests that can have an impact on students, instructors, and programs. This focus on the real world reflects the changes in language testing textbooks over the past three decades; the shift from providing basic information on assessment to demonstrating ways to integrate authenticity into assessment is striking. In addition, Carr's (2011) book includes a CD to help users apply the information in the text, with a specific emphasis on using statistics. Such approaches show that language testing texts are working to meet the needs of their users through contextualization and additional resources such as computer-based activities beyond a written text that allow users to practice what they have learned.

While the above provides only a glimpse into the kinds of text-based materials offered to classroom teachers, the very existence of such materials points to the

importance of assessment for language instructors, as well as an understanding on the part of textbook authors and publishers that theoretical texts were insufficient to explain testing to language instructors. It is also important to note that encyclopedias such as this one also provide a resource for language professionals to explore in depth a variety of issues in language assessment.

## Non-Text-Based Assessment Training

In addition to training provided by written texts employed during a formal university or graduate level class or independently, other formats have become available for training language instructors on assessment. This section outlines some self-paced self-instructional materials and web-based instructional materials for instructors.

### Self-Instructional Materials

Professional development workshops are frequent approaches to help instructors in all subjects supplement their formal training and improve their classroom effectiveness. With the proficiency movement in the United States in the 1980s, language instructors could participate (for a cost) in a 4-day training on oral proficiency assessment, a format previously restricted primarily to government employees.

As technologies became more accessible and less costly, tape-recorded materials, accompanied by tapes, could begin to replace live, face-to-face workshops; Kenyon and Stansfield (1993) and Kenyon (1997) investigated one new format: allowing potential language raters to participate in training through use of a kit rather than a live training workshop. Such self-instructional approaches allowed instructors to seek on their own (or upon advice from supervisors or other colleagues) new methods of language assessment to use in their classroom. Similarly, ETS developed self-training kits for raters of the SPEAK test; these kits included tapes and ancillary materials. These new formats allowed instructors who had not received training in new approaches during their education or for whom the approaches came after their formal education was completed to learn about and apply new testing methods.

As use of computers and the Internet grew throughout the 1990s, computer-based approaches gained in popularity throughout education. So, too, did access to more information on language assessment training.

Since 1995, Fulcher has hosted the *Resources in Language Testing* webpage (http://languagetesting.info, accessed 12/5/2015), which includes references, relevant organizations, and streaming video of well-known language testers responding to frequently asked questions in language testing on topics such as reliability, validity, test impact, item writing, and statistics. This page contains a plethora of useful information. Recently, he has added podcasts to accompany articles published in *Language Testing*, one of the two major journals devoted to language assessment. The addition of podcasts to supplement such academic articles demonstrates the growing need in academic journals, as in academic texts, of users to go beyond the written word and to use multiple forms of communication to describe and explain language testing to different users.

As the CEFR gains popularity in Europe, uses of it continue to grow. Among other useful resources is a "passport" to demonstrate student progress on the CEFR that students and instructors can complete to show student growth. These resources are available on the web and can be downloaded for use in schools. The Council of Europe has a website that provides resources on both the CEFR and assessment in general (http://www.coe.int/t/dg4/education/elp/, accessed 11/30/2015), including ways to develop an online portfolio to document language outcomes. The Centre for Canadian Language Benchmarks provides resources for learners and assessors on its website, including guidelines and resources for test development. Many European-based resources include information for language learners in addition to instructors; such resources are less plentiful for US-based resources. Two examples of learner-oriented resources in the United States are housed at the National Council of State Supervisors of Foreign Languages (NCSSFL) and CAL. NCSSFL developed a first paper-based and now online self-assessment system for US K-16 learners inspired by the CEFR efforts. This resource (http://www.ncssfl.org/LinguaFolio/index.php?linguafolio_index accessed 12/20/2015) is designed to help learners develop and track their progress toward language proficiency goals and requires registration. On a different note, in developing a new, computer-based Arabic oral proficiency assessment, CAL worked with learners to design a five-module online resource that describes different aspects of Arabic oral language proficiency, including both examples of student performances at different proficiency levels and clips of student interviews that describe how these students attained proficiency in Arabic (http://www.cal.org/aop/, accessed December 15, 2015).

In addition to resources for students, some organizations also provide support for teachers. In the late 1990s, the Center for Advanced Research on Language Acquisition (CARLA) of the University of Minnesota has developed a seven-module, online Virtual Assessment Center (VAC) to provide both resources, background information and guidance on second language classroom assessment (http://www.carla.umn.edu/assessment/vac/index.html, accessed 11/30/2015). The VAC includes an annotated bibliography of assessment resources, as well as a virtual item bank. The virtual item bank provides model items for teachers and is accompanied by item-writing tips. The VAC represents an early effort not only to help classroom language instructors develop good items and assessments for their students but also to understand the principles of assessment that undergird the process. Although the VAC is a valuable resource, the annotated bibliography has not been updated since the early 2000s. Perhaps one of the most challenging aspects of online resources is keeping them current; updating such resources regularly represents a significant commitment. If such resources are not reviewed regularly, they fall out of date quickly.

Swender et al. (2006) reported on a web-based survey of assessment uses and needs of 1,600 foreign language instructors in the United States. In addition to highlighting tests currently being used and needed for language instructors, the survey also highlighted a lack of understanding of many testing concepts, such as appropriate test use, by those who responded. As a result of this survey and other reports, in 2009, the Center for Applied Linguistics updated its foreign language test directory and developed a tutorial for users in test selection. In developing the

tutorial and soliciting feedback from a variety of stakeholders, Malone (2013) found a dichotomy between the perceived needs of such a tutorial by language instructors and by language testers. Language instructors stressed the needs for a succinct, understandable tutorial, while many language testing specialized and emphasized on the importance of explaining complex language testing concepts, such as assessment use and validity arguments, to such language instructors. The directory is updated biannually and the tutorial will be reviewed and updated by 2018. In 2015, the tutorial and directory received 63,000 unique views, thus highlighting the need for such online instruments.

## Works in Progress

Many of the current projects described are simultaneously works in progress and represent ongoing efforts to enhance both practice and understanding of assessment by language instructors. The addition of online tutorials, podcasts, videos, and e-portfolios across the world demonstrates the continued interest in and need for these resources. A recent edition of *Language Testing* was devoted to the issue of language assessment literacy; this special issue highlighted many facets of language assessment literacy from how language assessment is viewed in the parliament (Pill and Harding 2013) to the identity of the language tester (Jeong 2013) to the contrast between information valued by language testers and instructors (Malone 2013). In reviewing the wide range of topics addressed by this issue, it is clear that a variety of stakeholders could benefit from information about language assessment and that the audience for such information has expanded both beyond simply language testers and language teachers. As the field progresses, it is likely that still more online resources will become available; a likely issue to arise is how to evaluate the efficacy of the different resources to ensure that users not only use high-quality resources that reflect best practices but also that the resources they access are appropriate for their own needs. Although the university in general and teacher preparation programs in particular have been the traditional focus of language assessment, online resources represent an important way to provide both ongoing professional development to in-service teachers as well as basic information about language assessment to those outside the field. In addition to online resources, the International Language Testing Association (ILTA) provides funding for two or three workshops to be held annually in parts of the world where language assessment literacy could be improved or where such efforts are scarce.

## Problems and Difficulties

Although the landscape for including more stakeholders in the language assessment process and educating these stakeholders about language assessment is hopeful, it is nonetheless an ongoing and daunting task. The amount of resources available in print and online continues to grow, and a language instructor inexperienced in

language assessment might not understand how to select from among the many resources in the world.

This original chapter was released in 2008, and many more resources, from textbooks to online resources, have been released and are being used internationally. In 2008, the major challenge identified was determining who is and who should be trained in language assessment, how and to what extent such individuals are trained, and what the expected outcomes of such training should be. To some degree, raising the issue of training in language assessment is as important as any information contained therein; in 2008, the focus was on the training of language instructors. Although there is still no consensus on the assessment literacy needs of language instructors and, indeed, there are differences in perspective as to what language instructors believe, they need to know about assessment and what language testers believe that language instructors need to know (Malone 2013). It is important that the area of inquiry on language assessment literacy has expanded during the past 7 years beyond discussing what language instructors need to know. Pill and Harding (2013) investigated the assessment literacy needs from a parliamentary perspective; Jeong (2013) explored the language assessment literacy needs of testers and non-testers, and O'Loughlin (2013) explored the needs of university test users. However, one major gap that remains is how to best educate students about language assessment. Although three examples have been used in this chapter, the fact remains that students of language also need to understand why and how they are being assessed and how the results of their assessments will be used.

## Future Directions

As this chapter indicates, progress has been made to increase language assessment literacy efforts. The focus on language instructors and their understanding of assessment has expanded to include additional test users such as parents and administrators. The studies cited in this article represent the understanding that language assessment results are used for a number of far-reaching goals, from language students and teachers to our representatives in government. While this progress is helpful, additional work is needed. Pill and Harding (2013) mentioned above show that governments need education on assessment literacy; their efforts should be extrapolated to other governments. It is important to note, too, that students and test takers have not yet emerged as a focus of language assessment literacy research and this group is most affected by language tests and their results.

For continued progress to take place, it will be important to continue and expand work that explores both stakeholder perceptions of language assessments, the extent to which these perceptions are accurate, and how to mediate these expectations to help improve assessment literacy. As stakeholder language assessment literacy grows, it will become crucial for stakeholders to use this information to hold themselves and the developers of the tests they use accountable for the ways the tests are used and the decisions made on the basis of these tests. Finally, language learners themselves and their families must be included in this work. Learners and

their families benefit most and least from language assessment; thus, they must fully understand the tests they take and the implications the results have for them.

## Cross-References


- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ Language Assessment Literacy
- ▶ Washback, Impact, and Consequences Revisited


## Related Articles in the Encyclopedia of Language and Education

Nike Arnold: Technology and Second Language Teacher Professional Development, In Volume: Language, Education and Technology

Klaus Brandl: Task-Based Instruction and Teacher Training. In Volume: Second and Foreign Language Education

Linda von Hoene: The Professional Development of Foreign Language Instructors in Postsecondary Education. In Volume: Second and Foreign Language Education

Margaret E. Malone: Developing Instructor Proficiency in (Oral) Language Assessment. In Volume: Second and Foreign Language Education

Oleg Tarnopolsky: Nonnative Speaking Teachers of English as a Foreign Language. In Volume: Second and Foreign Language Education

## References


Alicea, I. (2005). NCLB requirements prompt changes in ELL assessment. *The ELL Outlook*, *4*(5), 19.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The modern language journal, 70*(4), 380–390.

Boyles, P. (2005). Assessment literacy. In M. Rosenbusch (Ed.), *National assessment summit papers* (pp. 11–15). Ames: Iowa State University.

Brindley, G. (1997). Assessment and the language Teacher: Trends and transitions. *Language Teacher* Online, 2.

Brown, J. D. (1998). *New ways of classroom assessment. New ways in TESOL series II*. Innovative Classroom Techniques. TESOL, Alexandria, VA.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Carroll, B. J. (1991). Resistance to change. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 22–27). London: Modern English Publications and the British Council.

Cohen, A. (1994). *Assessing language ability in the classroom*. Boston: Heinle and Heinle.

Davidson, F., & Lynch, B. (2002). *Testcraft*. New Haven: Yale University Press.

Davies, A., Brown, A, Elder, C., Hill, K., Lumley, T., & McNamra, T. (1999). *Dictionary of language testing*. Cambridge: University of Cambridge Local Examinations Syndicate and University of Cambridge.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education* (Cambridge Language Education). Cambridge/New York: Cambridge University Press.

Harris, D. P. (1969). *Testing english as a second language*. Washington, DC: ERIC Clearinghouse.

Hughes, A. (2004). *Testing for language teachers*. New York: Cambridge University Press.

Inbar-Lourie, O., Scarino, A., Malone, M. E., Jeong, H., O'Loughlin, K., Pill, J., & Harding, L. (2013). The special issue on language assessment literacy. *Language Testing, 30*(3), 301–307.

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers?. *Language Testing, 30*(3), 345–362.

Kenyon, D., & Stansfield, C. W. (1993). Evaluating the Efficacy of Rater Self-Training.

Kenyon, D. M. (1997). Further research on the efficacy of rater self-training. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 257–273). Jyväskylä: University of Jyväskylä.

Lantolf, J. P., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *The Modern Language Journal, 69*(4), 337–345.

Little, D. (2012). Elements of L2 proficiency: The CEFR's action-oriented approach and some of its implications. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the common European framework of reference for languages*. Tübingen: Stauffenburg Verlag.

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing, 28*(3), 367–382.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.

Moore, Z. (1994). The portfolio and testing culture. In C. Hancock (Ed.), *Teaching, testing and assessment: Making the connection* (pp. 163–182). Lincolnwood: National Textbook Company.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380.

O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Reading: Addison-Wesley Publishing.

Ommagio, A. (1986). *Teaching language in context*. USA: Heinlke & Heinle Publishers.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381–402.

Spolsky, B. (1977). Language testing: An art or science. In *Proceedings of the fourth international congress of applied linguistics* (pp. 7–28). Stuttgart: Hochsculverlag.

Spolsky, B. (1995). Measured words: The development of objetive language testing. In D. Kenyon, & C. W. Stansfield (1993) (Eds.), *Evaluating the efficacy of rater self-training*. Oxford University Press.

Stiggins, G. (1997). *Student centered classroom assessment*. Upple Saddle River: Prentice Hall.

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing: A resource for teachers and program administrators*. Alexandria: TESOL Publications.

Swender, E, Abbott, M, Vicars, R., & Malone, M. (2006). *The assessment of performance and proficiency in testing: a report*. Presentation at the convention of the American Council of Teachers of Foreign Languages, Nashville, TN.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Wiggins, G. (1994). Toward more authentic assessment of language performances. In C. Hancock (Ed.), *Teaching, testing and assessment: Making the connection* (pp. 69–86). Lincolnwood: National Textbook Company.

# Part III

# Assessment in Education

# Dynamic Assessment

Matthew E. Poehner, Kristin J. Davin, and James P. Lantolf

**Abstract**

Dynamic assessment, or DA, departs from the traditional distinction between formative and summative assessment, as it understands teaching to be an inherent part of all assessment regardless of purpose or context. This position follows from the theoretical basis of DA in the writings of Russian psychologist L. S. Vygotsky and in particular his proposal of the zone of proximal development. Positing that independent functioning indicates only abilities that have already fully developed, Vygotsky advocated procedures in which the assessor, or mediator, engages cooperatively with learners, offering support when learners encounter difficulties in order to determine the extent to which learners can extend their functioning as well as the forms of assistance to which they were most responsive. According to Vygotsky, this approach allows for a more in-depth diagnosis of learner development by revealing abilities that have not yet completed their development but are still emerging. In the decades since Vygotsky's death, his insight has generated a range of DA procedures undertaken with learners with special needs, immigrants, young children, and gifted learners as well as with individuals studying particular academic subjects, including second languages. L2 DA studies have generally been pursued in collaboration with classroom teachers, emphasizing dialogic

M.E. Poehner (✉)
World Languages Education and Applied Linguistics, Department of Curriculum and Instruction, The Pennsylvania State University, University Park, PA, USA
e-mail: mep158@psu.edu

K.J. Davin
Foreign and Second Language Education, School of Education, Loyola University Chicago, Chicago, IL, USA
e-mail: kdavin@luc.edu; kdavin@gmail.com

J.P. Lantolf
Language Acquisition and Applied Linguistics, Department of Applied Linguistics, Center for Language Acquisition, CALPER, The Pennsylvania State University, University Park, PA, USA
e-mail: jpl7@psu.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_18

interaction in one-to-one or small group settings. More recent projects have built upon this work to implement DA procedures in large-scale testing contexts. Current work is examining computerized administration procedures as well as uses of DA linked to curricular revisions intended to support learner appropriation of conceptual knowledge of language.

**Keywords**
Vygotsky • L2 development • Mediation • Zone of proximal development

## Contents



## Introduction

Lantolf and Poehner (2014) explain that a defining feature of sociocultural theory (SCT), as elaborated by L. S. Vygotsky (1987), is the central role it assigns to practical activity, especially education. In SCT, theory and research serve as an orienting basis for practice, which in turn provides the essential testing ground for theory, determining whether it should be accepted, revised, or rejected. According to Lantolf and Poehner (2014), this notion of "praxis" explains Vygotsky's keen interest in education, which he believed should aim to promote learner psychological development. The authors continue that this commitment to "developmental education" has guided much recent L2 SCT research, including work on dynamic assessment (henceforth, DA).

In DA, a teacher or assessor, referred to as a mediator, engages cooperatively with learners and intervenes when difficulties arise and their performance breaks down. Through a process of mediation, which is qualitatively different from corrective feedback, a diagnosis of learner development emerges that includes abilities that are fully formed, as indicated by learner independent performance, and abilities that are still emerging, determined by learner responsiveness during the mediating process. The activity of joint functioning with a mediator guides learners to perform beyond their current capabilities, thereby promoting their continued development. In this way, DA integrates teaching and assessing in a coherent framework. Since its introduction to the L2 field (Lantolf and Poehner 2004), DA has contributed to discussions concerning how classroom assessment may support student learning while also opening new directions in formal testing.

## Early Developments

Vygotsky's writings on the zone of proximal development (ZPD) provide the theoretical underpinnings of DA. The ZPD is based on the principle that higher forms of thinking (voluntary memory, attention, planning, learning, perception) are always mediated. Initially, they are mediated through our interactions with others and with physical and symbolic artifacts (e.g., books, computers, diagrams, language, etc.). These interactions are internalized and give rise to new cognitive functions. One's relationship with the world is still mediated, but this is accomplished on the internal plane of self-regulation. Consequently, Vygotsky (1998, p. 201) reasoned that assessments of *independent problem-solving* reveal only a part of a person's mental ability, namely, functions that have already fully developed. He termed this the *actual* level of development and contrasted it with the person's *potential* or *future* development, which he submitted could only be understood through their responsiveness during joint engagement with a mediator around tasks they are unable to complete independently.

An important corollary is that potential development varies independently of actual development, meaning that the latter, by itself, cannot be used to predict the former. This contrasts sharply with the belief in many approaches to assessment that a learner's future is more or less a linear continuation of the past, and hence the use of measures of independent performance on tests – reflecting the products of past development – to predict likely performance in the future. Vygotsky's discovery of the ZPD compels us to understand the future as not yet written but rather as resulting from continued access to appropriate forms of mediation, and its prediction is empirically based on learner responsiveness during cooperation with a mediator.

To our knowledge Vygotsky himself never used the term DA. The term may derive from his close colleague, A. R. Luria's (1961), description of ZPD assessments to differentiate children whose poor school performance resulted from biologically rooted disabilities, learning challenges, and language and culture differences. Critical to this diagnosis and to subsequent intervention planning was each child's responsiveness to mediation. Vygotsky and Luria's research laid the foundation for a range of formalized principles and procedures developed by researchers working with various populations around the world that have come to be known collectively by the name dynamic assessment (Haywood and Lidz 2007; Poehner 2008b). This work has been undertaken largely within special education and cognitive psychology and yielded a robust body of research dating from the 1960s.

In their review of DA research, Sternberg and Grigorenko (2002) note that the integration of mediation can be organized within the administration of an assessment or delivered as a distinct phase embedded between a pre- and post-administration of the test. They refer to these two models, respectively, as "cake" and "sandwich" formats. Representative of the cake format, Brown and Ferrara (1985) describe the use of mediation prompts and hints that are prescripted and arranged from most implicit to most explicit. The prompts are then offered to learners one at a time until the learner produces the desired response. An early example of the sandwich format

is Budoff's (1968) program that embedded a training module after the pretest to teach relevant principles. Sternberg and Grigorenko (2002) point out that both formats offer advantages: the sandwich format allows for comparisons between test performances prior to and following mediation, while the cake format streamlines the procedure and introduces mediation as soon as learners experience difficulties.

Lantolf and Poehner (2004) further differentiate DA models according to how mediation itself is conceived. They explain that much DA research, in both the sandwich and cake formats, limits mediation to a "one-size-fits-all" approach. By standardizing both the content of mediation, whether it be a training module or set of hints, and its delivery (i.e., provided in precisely the same manner to all learners), this work has aligned more closely with traditional testing practices and allowed greater use of inferential statistics for analyzing and comparing results. Lantolf and Poehner refer to these approaches to DA as "interventionist," highlighting that mediation is understood as prepackaged treatment. They point to another tradition in DA as "interactionist," and they suggest it more closely aligns with Vygotsky's understanding of cooperation in the ZPD. In interactionist DA, mediation follows the general principle of beginning in a more implicit manner and becoming increasingly explicit as determined by a learner's responsiveness to specific levels of mediation. Mediation is not scripted in advance but emerges through open dialogue with learners. This allows mediators considerable freedom to interact with learners, bring to the surface processes that underlie performance, and provoke further mediation (Poehner 2008b). According to Miller (2011), the mediated learning experience model of interactionist DA developed by Reuven Feuerstein (see Feuerstein et al. 2010) is a direct continuation of Vygotsky's and Luria's ZPD work. This research has been particularly influential in the development of L2 DA.

## Major Contributions

The first project to explore the use of DA in L2 education was undertaken by Poehner in his doctoral dissertation, which provided the basis for a book-length study (Poehner 2008b). This work details the theoretical origins of DA, overviews leading approaches, and documents the use of DA with university-level learners of L2 French. Two important contributions of that project are that it reconnected DA practices with Vygotsky's theory (a matter overlooked in much DA research outside the L2 field) and it provided detailed documentation of mediator-learner interactions, thus breaking with the convention in previous DA studies of reporting only outcomes of the procedures. Poehner's (2008b) analysis outlined particular moves on the part of the mediator and how they informed the diagnosis of learner development. As a follow-up, Poehner (2008a) examined the notion of "learner reciprocity," a concept that was proposed in earlier DA studies but for which there was little empirical data. Learner reciprocity refers to the range of behaviors learners may exhibit that go beyond correct or incorrect responses to mediation. Examples include eliciting mediator support, negotiating mediation, refusing offers of assistance,

posing additional questions, and seeking mediator approval. Together, the specific mediating moves and forms of learner reciprocity that characterize a DA session provide a nuanced profile of learner emerging abilities.

Close analysis of mediator and learner participation in DA, and the use of this information to interpret learner development, has been a consistent theme in L2 DA research. The major portion of this work has been conducted in instructional contexts, with the implementation of DA reflecting collaboration between researchers and teachers. The basis for much of this work has been Lantolf and Poehner's *Teacher's guide* to DA, now in its second edition (2011). The *Guide* includes a monograph that introduces DA principles and models, provides questions for discussion and resources for additional information, and walks readers through analysis of transcribed teacher-learner interactions showcasing the quality of mediation. The *Guide* also offers a series of video appendices illustrating examples of L2 DA. A *Casebook* of L2 DA studies (Poehner to appear) extends this with additional videos and analyses of collaborations with teachers that in fact emerged from previous workshops and uses of the *Guide*.

L2 DA has been pursued with learners at beginning through advanced levels of instruction, in primary school settings and universities, and with commonly taught languages such as Spanish and French as well as less commonly taught languages and even an indigenous Alaskan language for heritage speakers. Listening and reading comprehension, oral narrative abilities, pragmatic competence, and control over discrete grammatical features have each been a focus of mediation in DA research. A frequent question raised by the language teaching and assessment communities concerns the feasibility of moving beyond one-to-one interactions to include larger configurations of learners. In classroom settings, teachers are often responsible for groups of 20–30 learners, and sometimes more. In more formal assessment contexts, standardization is accepted practice in part because it allows large numbers of individuals to be assessed simultaneously. L2 researchers have begun to develop approaches to implementing DA principles under both these conditions.

Poehner (2009b) conceived of one approach to addressing numbers of learners in classroom settings by shifting the focus of mediation from the development of individuals to the group. Noting that Vygotsky (1998) himself raised the possibility of appropriately mediating a group ZPD, Poehner argues that DA in a group setting (G-DA) requires engaging learners in tasks that no individual can complete independently but that can be made accessible to every member of the group through appropriate mediation. In this way, there is both a struggle to stretch beyond one's current capabilities and a need for external forms of mediation. Poehner (2009b) discerns at least two forms of G-DA. "Concurrent" G-DA occurs as a mediator that engages a group or an entire class in an activity and negotiates mediation with the group. Pointing to an analysis of classroom interaction reported by Gibbons (2003) involving ESL learners working to appropriate scientific discourse, Poehner notes that in concurrent DA the mediator may address particular individuals, providing prompts to one, leading questions to another, and so on. The specific mediating behaviors directed at an individual are not the focus, however, as it is the interaction in its entirety that provides insights into the understandings and abilities of the group.

In concurrent G-DA, given variability across learners, not every mediating move will be relevant to each individual. Some will move more quickly toward independent performance than others. We will have more to say about this later, but for now we point out that the matter is at least partly addressed in "cumulative" G-DA. Here, interactions unfold between the mediator and individual learners one at a time, and on the face of it, this approach appears to be a one-to-one administration. The crucial difference, however, is that the interactions occur in a class setting, with the expectation that other learners are engaged as "secondary interactants." In other words, even though the rest of the class may remain silent while the mediator engages with an individual, the interaction itself has the potential to mediate each learner's thinking. This approach to G-DA thus aims for a cumulative effect of mediation wherein learners who work with the teacher later in a lesson may also reference the mediational processes from previous exchanges in the class. Indeed, Poehner (2009b) offers an example of cumulative G-DA from an L2 Spanish elementary school class. His analysis of three learners who each take a turn participating in a game in the L2 reveals a steady reduction in the degree of teacher mediation required as the game progresses. Poehner suggests that in reality, the second and third learners may have already benefitted from mediation prior to the start of their turn.

With regard to large-scale testing, Guthke and Beckmann (2000) recognized the potential of increasingly sophisticated computer programs to assume the role of mediator. Mediation made available in a computerized DA (C-DA) administration certainly does not allow for the careful alignment with learner need characteristic of interactionist DA. Nonetheless, it offers the possibility to move beyond ascertaining the correctness of a learner's response and indicates if s/he is able to reach the solution when mediation is offered. Guthke and Beckmann describe a tutorial approach developed for use with a C-DA version of their *Leipzig Lerntest*, a cognitive aptitude instrument. Although the authors do not provide specific examples from the test or data from its administration, they explain the principle as suspending the test when a learner incorrectly answers a question in order to introduce a brief tutorial that explains relevant principles and walks learners through practice problems. Once the tutorial ends, the test resumes and the learner is presented with a parallel version of the item she/he had missed. In this way, it is possible to distinguish learners who answered questions correctly without intervention, those whose performance improved following the tutorial, and those whose difficulties persisted in spite of the available mediation. The authors maintain that this more nuanced diagnostic of learner abilities is helpful to designing remediation programs specific to learner needs.

Poehner and Lantolf (2013) see a similar potential for C-DA in the L2 domain, underscoring a diagnosis that takes account of learner emerging abilities as having immediate relevance for placing learners at appropriate levels of study in language programs. They designed C-DA tests of listening and reading comprehension across three languages: Chinese, French, and Russian. The tests were modeled after existing standardized measures of L2 comprehension and followed a multiple-choice format. They departed from the convention of providing four options for each test

item (the correct answer and three distractors), preferring instead to add a fourth distractor. This increased the number of times an examinee could attempt the items and the number of mediating prompts that could be offered. Following Brown and Ferrara's (1985) graduated-prompt approach, the program generates two scores: an "actual" score, reflecting whether an examinee's first response was correct, and a "mediated" score calculated to indicate the number of attempts an individual made – and, hence, the number of mediating prompts required – in responding to a test item (Poehner and Lantolf 2013). The logic of this approach was that a learner who answered on, for instance, a second attempt was likely to have better comprehension of a text than a learner who required three or four attempts or who was not able to reach the correct answer even after all four mediating prompts were provided. In addition, an explanation in English was offered to learners after the item was correctly answered and before the next item was presented. Thus, learners had access to learning opportunities during the test itself, an important feature of DA.

The C-DA tests are available online and are cost-free (www.calper.la.psu.edu). Analyses of scores generated by the tests provide evidence in support of Vygotsky's prediction that learner mediated performance varies from independent performance in ways that cannot be determined a priori. In the context of the L2 comprehension tests, this means that actual scores are not always indicative of mediated scores; therefore learners with the same actual score may have different mediated scores reflecting different degrees of prompting. An attractive feature of the C-DA tests is that items are grouped according to the underlying construct (within listening or reading comprehension) and a profile is automatically created by the program for individual learners. This allows one to observe learner performance in specific areas of language ability, such as the lexicon and sentence-level or discourse-level grammar. In addition to informing placement decisions, learner profiles are useful for classroom teachers in shaping instruction to the needs to individual learners or groups of learners.

## Work in Progress

DA has stimulated interest across a range of different areas of L2 research. In this section we limit our discussion to three areas that we believe will continue to be important for the future of DA. The first builds upon the concept of G-DA to bring DA into day-to-day classroom activities. As an example of this work, we consider one of the "cases" documented in the DA *Casebook* (Poehner to appear) that documents a teacher's effort to reorganize her advanced level L2 Japanese composition course.

Originally designed according to a "process approach" to writing, the course required learners to produce multiple drafts of their work, which they shared and revised through the following stages: a one-to-one writing conference with the teacher, a peer-editing session in class that involved students working in pairs or small groups to read and comment upon one another's work, and whole-class discussions of advanced features of Japanese grammar. As Poehner (to appear)

explains, the teacher, Sayuri, undertook to revise each of these writing stages according to how the students responded to mediation. One-to-one writing conferences were refashioned as interactionist DA sessions in which initial drafts were reviewed and specific language problems were identified. These individualized sessions allowed Sayuri to identify which features of Japanese were within learners' emerging ability to control the language, determined by their responsiveness to mediation. Learners were then placed into groups of two or three based on similar sources of difficulty and given a packet of sentences containing errors drawn from their compositions. In this way, the more traditional peer-editing step in process writing became focused on problems that were within the ZPD of each member of the groups. After the students reviewed the sentences, made corrections, and prepared explanations of their proposed revisions – an activity intended to prompt learners to support one another's understanding of relevant features of the L2 – Sayuri reviewed the packets with the entire class. This final stage of the approach represented a larger G-DA and served to clarify misunderstandings, discuss alternative corrections, and make connections across similar types of learner problems.

As analysis of data from this project continues, particular attention is being given to the quality of learner interaction during the G-DA peer review as well as the kinds of contributions made by the teacher during the larger G-DA context. The latter is of interest because it differentiates between problems that were appropriately resolved during peer work and those that needed further mediation from the teacher. With regard to the former, it would seem plausible that by grouping learners according to their ZPD, it biases them in favor of working cooperatively to revise their papers. Whether this occurred and promoted the development of all learners is a crucial question that is yet to be resolved.

Another area of interest concerns the teacher's experience with DA. The focus here is on the preparedness of L2 teachers to deploy SCT principles along the lines of the developmental education argued for by Lantolf and Poehner (2014). The *Casebook* includes interviews with teachers reflecting on their understanding of DA, the reasons behind their decision to integrate it into their practice in a particular manner, and the challenges they may have experienced in so doing. Analysis of the interviews is currently underway, and it is anticipated that the information will provide a resource for teachers and researchers to better understand the demands of implementing DA and how these might be addressed.

Davin and Herazo (2015) are investigating how teachers' experiences with DA may raise their awareness of the discursive practices that characterize their interactions with learners, which the authors consider to be an essential step toward creating classroom discourse patterns to promote learner agency. The participants, which include in-service English teachers in Montería, Colombia, and preservice Spanish and Italian teachers in Illinois, USA, studied the DA *Guide* (Lantolf and Poehner 2011) and participated in professional development seminars to support DA implementation. Using a qualitative case study design, Davin and Herazo compare the participants' pre-DA and post-DA classroom discourse patterns. Preliminary analysis of lesson transcripts and stimulated recall sessions suggest that DA prompted more dialogic classrooms, fostering an environment characterized by a more equal

balance of teacher and student talk and extended interaction sequences between the teacher and learners.

A third area receiving attention from researchers concerns applications of DA in large-scale testing contexts. Levi (2012) suggests that because DA creates possibilities to promote learner development, it functions to produce a kind of positive washback wherein an existing formal testing program becomes not only a means of measuring learner abilities, but it may also provide an opportunity for learning to occur that complements learning opportunities already present in classrooms. Working within the context of large-scale oral proficiency interviews among secondary school students in Israel, Levi (2012) constructed mediating resources around the rubrics employed to assess dimensions of language proficiency, including fluency and accuracy. She then designed a DA procedure following the sandwich format described by Sternberg and Grigorenko (2002) and added a fourth step: a delayed posttest, or transfer test, intended to determine the durability of any gains made by learners.

Levi (to appear) reports a study using this procedure in which she recruited a total of 73 Israeli secondary students and divided them into three groups: a control group, which received no mediation between the pre- and posttests, and two mediation groups. The two mediation groups were further differentiated according to whether learners worked independently or as part of a group. In both cases, mediation occurred across four sessions and included the presence of a tester mediator to facilitate learner engagement in the activities and their use of the assessment rubric. In the first session, learners reviewed a recording of either their own pretest performance or that of another participant. They worked to apply the rubric to an evaluation of the performance, which positioned them for interactions in the subsequent sessions as they attempted to use the rubrics to monitor their own speaking practice. Levi (to appear) reports that students in both the mediation groups improved their posttest performance, while those in the control group actually scored lower. More modest gains on the transfer assessment were also found for students who had received mediation. This research offers compelling evidence that indeed DA can be "scaled up" to function in large-scale testing situations and that this may be done in a manner that preserves DA's commitment to both diagnosing and promoting learner development.

## Problems and Difficulties

As explained, a challenge for DA has been moving beyond one-to-one contexts of the sort documented in Poehner's (2008b) initial exploration of the framework. C-DA and G-DA offer viable ways forward, and we encourage additional work in both these areas. That said, one critique commonly leveled against DA is that it merely represents "good teaching" and nothing more. We concur that DA does indeed constitute effective teaching, but we further insist that effective instruction necessarily entails effective assessment – assessment with a future rather than a past orientation. In other words, assessment that promotes learner development. The L2 research literature is replete with contradictory findings and recommendations to

teachers concerning implicit forms of feedback such as recasts or explicit corrective feedback. Likewise, research on formative assessment has long found that teachers are likely either to emphasize affective support and encouragement at the cost of helpful feedback or to over- or underestimate learner abilities (Torrance and Pryor 1998).

Our experience collaborating with teachers suggests that prior to learning about and experimenting with DA, it is highly unlikely that they systematically provide appropriate mediation to learners. While there is variability concerning how sensitive teachers are to learner needs, without a coherent theory to guide their actions, mediation is either offered in a hit-or-miss manner, sometimes attuned to learner responsiveness but not always, or it is provided in a one-size-fits-all approach in order to treat all learners the same (see Lantolf and Poehner 2013). It often requires considerable effort to help teachers move toward interactions that take account of changes in learner needs and responsiveness during joint activity. Indeed, the classroom teacher in Poehner's (2009b) study preferred an interventionist approach to DA even though she was not using it for a formal assessment purpose; standardization was appealing precisely because it mitigated the demands of an open-ended procedure. That said, both the *Guide* and the *Casebook* offer examples of impressive creativity and thoughtfulness on the part of teachers in implementing DA once they have come to understand its principles and theoretical foundation.

Another critique of DA stems from the fact that it does not adhere to accepted testing practices, in particular standardization of procedures. This concern seems less relevant to instances of C-DA or interventionist DA more generally, which as explained commit to standardization with regard to mediation and the interpretation of results. Nonetheless, the fact that DA departs from conventions of standardized testing has been a concern since before its introduction to the L2 field. For instance, Büchel and Scharnhorst (1993, p. 101) concluded that DA could not be taken seriously until it committed to measurement, which they proposed demands "standardization of the examiner-subject interaction." Glutting and McDermott (1990, p. 300) similarly criticized the "creative latitude" in approaches to DA such as Feuerstein's because some learners receive more help than others. Within the L2 field, this line of criticism is echoed by Fulcher (2010, p. 75), who expresses the view that because mediator and learner function jointly insights from DA cannot be generalized beyond a particular "instance of occurrence" involving the given task and participants. Moreover, he faults DA for not taking account of how the presence, absence, or strength of particular factors can yield testable predictions of learner development.

Lantolf and Poehner (2014) respond to Fulcher's critiques in detail, including claims he makes about SCT in general. We will not repeat those remarks here, but we do wish to point out that Poehner (2007) dealt extensively with the topic of generalizability. As he explained, research in both interventionist and interactionist traditions frequently present learners with tasks that are either designed to employ the same underlying principles as those used throughout the assessment but in new combinations or applied to more difficult problems. The point of requiring learners to extend their performance beyond a given set of tasks, a practice alternately

referred to as "transcendence" or "transfer," is to ensure that the effects of mediation are not task specific, limited to the here and now, but rather that they represent actual change in psychological functioning. Recall that the purpose of DA is not to help learners do better on a given assessment task, which distinguishes DA from scaffolding (see Lantolf and Poehner 2004), but to promote their development, that is, to generalize the mediation they have appropriated in a given task and context to new tasks and contexts. For this reason, the different forms of mediation and how learners respond (the presence, absence, and strength of variables) are given much attention in DA and are typically traced over time.

## Future Directions

Poehner (2009a) argues that the full potential of DA to promote learner development might be realized through a two-pronged approach in which the same principles of mediation guide both formal assessments and classroom activities. Following from the discussion of L2 C-DA, formal evaluation of learner abilities that takes account of the ZPD (i.e., their emerging abilities and the future investment likely required before they reach independent functioning) will in some cases lead to different decisions regarding acceptance of learners into programs and placement at an appropriate level of study. An important topic for future research will be to empirically investigate ZPD-based predictions of learner development. This research would entail following learners longitudinally to document development over the course of L2 study and how their progress reflects their DA performance. Of course, realizing their potential is dependent upon continued mediation that is sensitive to their emerging abilities and that changes in step with their development. In other words, the instruction itself must be of the sort that aims to promote learner abilities in the L2. It is here that two intersecting lines of research can be carried out in tandem with DA: systemic-theoretical instruction (STI) and mediated development (MD).

Briefly, STI compels a reorganization of L2 curricula and indeed a refocusing of the goals of L2 instruction. Based on Vygotsky's analysis of the value of teaching that brings abstract theoretical knowledge in contact with learners' practical experiences, STI shifts away from traditional form-focused L2 instruction in favor of instruction grounded in conceptual knowledge of the language. Following Vygotsky, abstract conceptual knowledge goes beyond what learners would likely "figure out" for themselves from everyday experiences in the world. Moreover, STI presents concepts in a systematic manner that avoids problems associated with discovery learning (Karpov 2014). The goal of STI is to help learners develop understandings of the central concepts in a field of study, how these concepts interrelate, and how together they provide an appropriate orienting basis for action. L2 STI studies to date have targeted topics such as interactional pragmatics in French, sarcasm in English, and topicalization in Chinese (Lantolf and Poehner 2014). Internalization of L2 conceptual knowledge allows learners to use the language in intentional ways that break from concerns over prescriptive rules and to instead understand language as a resource for the creative, nuanced formation and expression of meanings. The

diagnosis of development that emerges from DA affords crucial insights for understanding and guiding learner progress through an STI program. Specifically, DA reveals learner understandings that are behind their use of language, the extent to which they have begun to internalize conceptual knowledge, and specific forms of mediation that promote their use of the concepts during communicative activity.

Related to the integration of DA and STI, Poehner and Infante (2015) propose that mediator-learner cooperation may shift from a focus on diagnosing learner abilities in favor of more strongly emphasizing the teaching component of the interaction. This does not undermine the relation between assessing and teaching as two features of ZPD activity, that is, the activity of understanding and promoting development. Rather, the point is that in any instance of mediator-learner cooperation, one may bring to the fore either the assessing or teaching function so long as one does not lose sight of the other. Selecting a focus requires planning on the part of the mediator to determine the goal of a particular interaction. Drawing on a project that included mediator-learner interaction throughout an STI program for L2 English learners, Poehner and Infante (2015) report that cooperative interaction proved essential for introducing conceptual knowledge to learners, presenting specialized instructional materials associated with STI (e.g., models, charts, and images), modeling how these resources function as tools for thinking, and supporting learner efforts to integrate the concepts into their meaning making in the L2. The authors propose the term MD for such interactions to underscore the focus on teaching to promote development. In Poehner and Infante's analysis, this shift in focus manifested in changes in mediator contributions, specifically with less effort to provide prompts and leading questions to learners and an increase in explanation and verbalization of the mediator's understanding of the materials and their relevance for orienting to activity and reflecting on outcomes. To be sure, this is only an initial exploration of MD. More work is needed to understand the forms that mediation may take in such interactions, how they overlap with DA, and how the alternating foci of assessing and teaching function together to guide learner development.

## Cross-References


- ▶ Language Assessment Literacy
- ▶ Task and Performance-Based Assessment
- ▶ Using Portfolios for Assessment/Alternative Assessment


## Related Articles in the Encyclopedia of Language and Education

Rémi A. van Compernolle: Sociocultural Approaches to Technology Use in Language Education. In Volume: Language, Education and Technology

Amy Ohta: Sociocultural Theory and Second/Foreign Language Education. In Volume: Second and Foreign Language Education

Rebecca L. Oxford: Conditions for Second Language (L2) Learning. In Volume: Second and Foreign Language Education

## References


Brown, A., & Ferrara, R. A. (1985). Diagnosing "zones of proximal development". In J. V. Wertsch (Ed.), *Culture, communication and cognition. Vygotskian perspectives*. Cambridge: Cambridge University Press.

Büchel, F. P., & Scharnhorst, U. (1993). The learning potential assessment device (LPAD): Discussion of theoretical and methodological problems. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues*. Amsterdam: Swets & Zeitlinger.

Budoff, M. (1968). Learning potential as a supplementary testing procedure. In J. Hellmuth (Ed.), *Learning disorders* (Vol. 3). Seattle: Special Child.

Davin, K. J., & Herazo, J. D. (2015, March). *Transforming classroom discourse through dynamic assessment*. Paper presented at the annual meeting of the American Association of Applied Linguistics, Toronto.

Feuerstein, R., Feuerstein, R. S., & Falik, L. H. (2010). *Beyond smarter: Mediated learning and the brain's capacity for change*. New York: Teachers College, Columbia University.

Fulcher, G. (2010). *Practical language teaching*. London: Hodder Education.

Gibbons, P. (2003). Mediating language learning: Teacher interactions with ESL students in a content-based classroom. *TESOL Quarterly, 37*, 247–273.

Glutting, J. J., & McDermott, P. A. (1990). Principles and problems in learning potential. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children. Intelligence and achievement*. New York: Guilford.

Guthke, J., & Beckmann, J. F. (2000). The learning test concept and its applications in practice. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications*. Amsterdam: Elsevier.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice. Clinical and educational applications*. New York: Cambridge University Press.

Karpov, Y. V. (2014). *Vygotsky for educators*. New York: Cambridge University Press.

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics, 1*, 49–74.

Lantolf, J. P., & Poehner, M. E. (2011). *Dynamic assessment in the foreign language classroom: A teachers' guide* (2nd ed.). University Park: Center for Advanced Language Proficiency Education and Research, The Pennsylvania State University.

Lantolf, J. P., & Poehner, M. E. (2013). The unfairness of equal treatment: Objectivity in L2 testing and Dynamic Assessment. *Educational Research and Evaluation, 19*, 141–157.

Lantolf, J. P., & Poehner, M. E. (2014). *Sociocultural theory and the pedagogical imperative in L2 education. Vygotskian praxis and the research/practice divide*. London: Routledge.

Levi, T. (2012). *The effect of Dynamic Assessment on the performance of students in oral proficiency tests in English as a foreign language*. Unpublished doctoral dissertation, Tel Aviv University, Tel Aviv.

Levi, T. (to appear). Developing L2 oral language proficiency using concept-based Dynamic Assessment within a large-scale testing context. *Language and Sociocultural Theory.*

Luria, A. R. (1961). Study of the abnormal child. *American Journal of Orthopsychiatry. A Journal of Human Behavior, 31*, 1–16.

Miller, R. (2011). *Vygotsky in perspective*. New York: Cambridge University Press.

Poehner, M. E. (2007). Beyond the test: L2 Dynamic Assessment and the transcendence of mediated learning. *The Modern Language Journal, 91*, 323–340.

Poehner, M. E. (2008a). Both sides of the conversation: The interplay between mediation and learner reciprocity in Dynamic Assessment. In J. P. Lantolf & M. E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 33–56). London: Equinox Publishing.

Poehner, M. E. (2008b). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. Berlin: Springer.

Poehner, M. E. (2009a). Dynamic Assessment as a dialectic framework for classroom activity: Evidence from second language (L2) learners. *Journal of Cognitive Education and Psychology, 8*, 252–268.

Poehner, M. E. (2009b). Group Dynamic Assessment: Mediation for the L2 classroom. *TESOL Quarterly, 43*, 471–491.

Poehner, M. E. (to appear). *A casebook of Dynamic Assessment in foreign language education*. University Park: Center for Advanced Language Proficiency Education and Research, The Pennsylvania State University.

Poehner, M. E., & Infante, P. (2015). Mediated development as inter-psychological activity for L2 education. *Language and Sociocultural Theory, 2*, 161–183.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized Dynamic Assessment. *Language Teaching Research, 17*, 323–342.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing. The nature and measurement of learning potential*. Cambridge: Cambridge University Press.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom*. Buckingham: Open University Press.

Vygotsky, L. S. (1987). In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L. S. Vygotsky. Volume 1, Problems of general psychology, including the volume thinking and speech*. New York: Plenum.

Vygotsky, L. S. (1998). In R. W. Rieber (Ed.), *The collected works of L. S. Vygotsky. Volume 5, Child psychology*. New York: Plenum.

# Language Assessment Literacy

Ofra Inbar-Lourie

**Abstract**

Language assessment literacy (LAL) refers to the knowledge skills and principles that stakeholders involved in assessment activities are required to master in order to perform assessment tasks. The need for defining a literacy framework in language assessment has arisen following acknowledgment of teachers' assessment needs as well as the increase in the number of stakeholders from different disciplines involved in language assessment activities and decision-making. Though gaining momentum in theory and practice, the conceptual LAL framework is still in an evolutionary phase, with central unresolved issues. One of the main issues is the gap between formative or dynamic assessment perspectives and a focus on testing expertise. Attempts to define the LAL canon, specifically the role of language features within that canon, are still undecided. Promising endeavors have recently been made at operationalizing the theoretical framework in order to allow for the design and implementation of LAL initiatives. Future directions point at a move towards situated differential rather than unified LAL conceptualization in the form of language assessment literacies.

**Keywords**

Language assessment literacy • Assessment literacy • Language assessment culture • Assessment for learning

O. Inbar-Lourie (✉)
The School of Education, The Program for Multilingual Education, Tel Aviv University, Tel Aviv, Israel
e-mail: ofrain@tauex.tau.ac.il; ofrain@post.tau.ac.il

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_19

## Contents



## Introduction

The concept of language assessment literacy (henceforth LAL), which draws from general assessment literacy (henceforth AL), refers to the knowledge stakeholders need in order to conduct language assessment activities (Fulcher 2012; Taylor 2013). Though the concept in its present form is relatively new, it is gradually generating a growing body of literature and research. Studies focus primarily on teachers and their assessment knowhow, but also on other professionals in related areas in need of assessment expertise (e.g., O'Loughlin 2013). Like AL, which emerged amidst discussion of the transition from testing to assessment cultures (Shepard 2000), LAL writings also suggest adopting a constructivist sociocultural approach to language learning and assessment in light of the social turn in language testing (McNamara and Roever 2006) and the debate over the responsibility of the language tester (ILTA Code of Ethics 2000).

Attempts to define the LAL framework are in progress amidst a debate over central controversial issues. These include examination of an agreed upon canon of language testing knowledge as well as decisions as to who should disseminate this knowledge and accompanying skills (Heejeong 2013). A primary crucial issue revolves around the role of the language component within language testing knowledge, i.e., in what ways LAL differs from general assessment literacy and hence merits recognition in its own right (Inbar-Lourie 2013). The debate over what constitutes LAL also brings to the fore discussion on the relationship between language testing and other disciplines (Davies 2008) as well as the differential assessment knowledge requirements of stakeholders further away from the "assessment core" (Taylor 2013). However, most importantly, the LAL debate can be seen as an attempt to explore and define the language testing profession vis-à-vis internal and external mitigating factors, especially the match between the dynamic nature of language development and use on one hand and language assessment practices on the other. The review will first look at general assessment literacy and then focus on the specific theory and traits of language assessment literacy arguing for recognition as an entity in its own right.

## Early Developments

The origins of language assessment literacy lie in the emergence of the term "assessment literacy" in general education (Stiggins 1991), defined as "an individual's understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (Popham 2011, p. 267). The concept made its debut following acknowledgment of the central role that teachers play in the assessment process. Teachers are viewed as both consumers of testing information and independent assessors, hence requiring "the knowledge of means for assessing what students know and can do, how to interpret the results from these assessments, and how to apply these results to improve student learning and program effectiveness" (Webb 2002, p. 1). The contents of AL have been shaped largely by the growing emphasis on the formative role of assessment (Black and Wiliam 1998), particularly on the need to provide constructive feedback to advance learning.

Though discussion of AL has in recent years shifted to include additional protagonists involved in assessment activities, many of the writings and studies in this area still focus on practitioners and the impact of their assessment knowledge and activities on their students. An important landmark in defining teachers' assessment literacy is the publication of the *Standards for Teacher Competence in Educational Assessment of Students* (The American Federation of Teachers et al. 1990). The standards delineate the skills teachers need in seven domains: in choosing and developing assessment methods; in administering, scoring, and interpreting assessment results; in using assessment results for decision-making and grading; in communicating assessment results; and in recognizing unethical, inappropriate assessment use, and information. Research findings on teachers' knowledge in the areas specified in the standards are rather grim, pointing at the inadequate knowledge of both pre- and in-service teachers (Mertler 2002; DeLuca 2012).

Language assessment literacy (LAL) has drawn considerably from the literature and research on general assessment literacy (AL), while attempting to set itself apart as a knowledge base that incorporates unique aspects inherent in theorizing and assessing language-related performance. Yet, as Davies asks (in an article accompanying the Dictionary of Language Testing, Davies, Brown, Elder, Hill, Lumley and McNamara 1999), "how does one determine what count as fundamental concepts informing constructs of language testing?" (p. 244). Studies attempting to answer this question have referred primarily to two sources intended to disseminate language testing knowledge: language testing textbooks and language testing courses. A review of the former (Davies 2008) shows that while in the past textbook writers were concerned with "a knowledge + skills" approach to language testing (skills referring to "the practical know-how in test analysis and construction" and "knowledge" of the "relevant background in measurement and language description," Davies, p. 328), an additional component referred to as "principles" has been added. Principles are defined as "the proper use of language tests, their fairness and impact, including questions of ethics and professionalism" (ibid, p. 335).

With regard to language testing courses, findings of a study on the contents of language testing courses conducted in 1996 by Bailey and Brown and replicated 12 years later (Brown and Bailey 2008), reaffirm Davies' "knowledge + skills" categorization. Language testing experts participating in the study were asked to indicate the extent to which certain topics (presented as items on a questionnaire) formed part of their testing courses. The items dealt almost exclusively with tests and their properties, excluding references to the larger assessment picture – to assessment considerations and consequences as well as alternative forms of assessment. Testing culture was thus established as the core canon of language testing knowledge. Some of the agreed upon topics amongst the respondents were test critiquing and test analysis, item writing for the different skills, item quality and discrimination, validity, reliability, and standard error of measurement. The emerging knowledge framework can hence be classified as oriented predominantly to educational measurement rather than to language learning, as hardly any mention was made of language-related issues. Overall, results for both the 1996 and 2008 studies were generally similar, thus indicating (according to the researchers) the existence of a stable language testing knowledge base.

The social and critical turn in language testing (McNamara and Roever 2006; Shohamy 2001) signaled a shift away from a testing-oriented LAL focus to awareness of the need to include a dialogical assessment culture (Inbar-Lourie 2008a), one which fosters contextually relevant and diverse assessment practices while also paying heed to the "principles" category referred to by Davies (2008). Some of the concepts introduced at this point were already part of the aforementioned Standards for Teacher Competence in Educational Assessment of Students (1990), particularly the use of diverse assessment instruments to fit different purposes but also the standards of tailored assessment, the consequences of assessment, and the need for an ethical code in the assessment process. What made this shift more meaningful was that it was now being applied not only to contemporary language-specific dilemmas and challenges but also to the individuals impacted by assessment procedures and decisions, for example, adult immigrants and second language learners in the school context (Barni 2015; Elder 2015).

In addition, multilingual realities, alongside the spread of English as a Lingua Franca, have presented new testing challenges and the need to consider different assessment practices as well and identify ensuing assessment literacies. The concept of multilingual testing, whereby test takers are offered multilingual assessment tools (Shohamy 2011), reflects this change. Significant developments in language teaching pedagogy over the last 15 years require new assessment modes. Translanguaging, for example (García and Wei 2014), an approach which views meaning making as a holistic hybrid process that transcends language borders, requires matching assessment modification and knowhow. New assessment considerations have also arisen with regard to the growing use of the target or additional language as a medium of instruction in Content Language Integrated Learning (CLIL) or English Medium of Instruction (EMI) models, in both K-12 and tertiary institutions, "with no simple solutions at hand" (van Leeuwen 2006, p. 20). Moreover, the emergence and growing influence of a major player in the language

teaching and assessment scene in this third millennium, the Common European Framework of Reference (Council of Europe 2001), has brought into discussion the integration between the teaching learning process and assessment. Alongside proficiency standards, the CEFR has also introduced and promoted alternative forms of assessment, such as the European Language Portfolio and self-assessment. The move towards integrating teaching and assessment is also evident in the "learning-oriented assessment" approach (Purpura and Turner 2013) and in sociocultural approaches to language learning and evaluation, specifically the implementation of Vygotskian dynamic assessment concepts (Poehner 2009).

These developments, in tandem with the gradual permeation of formative assessment for learning considerations, are resulting in a slow transition in the LAL discourse towards a more expanded conceptual and practical repertoire. The evidence of this rather nascent transition seems to be rather confounding for the language testing community, for its roots are clearly in the testing tradition rather than in the implementation of a more comprehensive assessment framework that is integrated with learning and includes a variety of assessment tools.

One of the first attempts to introduce an assessment literacy framework within language testing can be accredited to Brindley and his modular framework (2001). The framework was intended for teachers and comprises of both core and optional elements, acknowledging the differential needs of language teachers. The modules reflect a wide perspective in an attempt to address some of the issues and dilemmas that have arisen in language assessment. The first module (perceived as core) provided the background to assessment from social, educational, and political perspectives, while the second core module, "Defining and describing proficiency," related language assessment to language knowledge models and looked at issues of validity and reliability. The next two modules were optional, focusing on language tests as well as on a more curricular classroom-embedded orientation. The last optional module presented a more advanced discussion of language assessment and research intended for teachers planning test construction projects or assessment-related research (Brindley 2001, pp. 129–130). Hence, the suggested LAL framework was compiled to match the nature of language knowledge and the resultant assessment literacy required.

Understandings, however, as to what is vital for becoming literate in language assessment and the depth of the knowledge needed are seen to fluctuate depending on the stakeholders involved and/or on the assessment context. Malone (2013) found that when language testers were asked about preferred contents in a language testing professionalization initiative they tended to focus more on the theoretical aspects, while language teachers opted for allocating greater weight to assessment tasks. Moreover, research examining the perceptions of language testing instructors who are specialists in the field versus applied linguistics nonlanguage testing specialists points at differences in terms of the respective perceptions of each group as to the topics that should be included in language testing courses (Heejeong 2013).

Attempts have been made over the last few years to relate to these different protagonists and describe and define their LAL needs and understandings. The following section will present the major contributions in this area referring first to

language teachers, who still comprise the primary LAL target population, moving on to additional constituents.

## Major Contributions

The predominant teacher-focus evident in general AL writings is also prominent in the LAL literature, coinciding with an overall interest in language teachers as assessors (Davison and Leung 2009). Fulcher (2012) notes the need for teacher assessment literacy: "If language teachers are to understand the forces that impact upon the institutions for which they work and their daily teaching practices, and to have a measure of control over the effects that these have, it is important for them to develop their assessment literacy" (pp. 114–115).

However, unlike the teacher standards assessment framework (1990), no equivalent document that describes the particular knowledge language teachers are required to have in order to perform assessment duties has thus far been offered. This may be due to the paucity of data available till recently on language teachers' LAL needs, but it may also be reflective of the theoretical transition and uncertainties evident in the field.

The European Association of Language Testing and Assessment (EALTA), upon its establishment (2004), undertook to hold language assessment training activities, noting however the need to identify first the existing knowledge of the various target audiences. The ensuing survey conducted for that purpose (Hasselgreen et al. 2004) included 914 respondents in the European context divided into three groups – teachers, teacher trainers, and experts, with teachers forming the majority of the sample. Findings showed that teachers and their trainers attested to lacking training in "the less traditional areas of assessment, such as portfolios, including the European Language Portfolio, and testing to the CEF." Similar teachers' assessment needs surveys have been conducted over the last decade in different geographical locations representing these different audiences, all attempting to dispel the LAL mist, recommending and/or designing professionalization initiatives. These studies have had a vital role in surfacing LAL issues and alerting the profession of the unattended building blocks required for establishing literacy in the field (e.g., Huhta et al. 2005; Vogt and Tsagari 2014).

Interestingly, unlike the general AL surveys conducted among pre and in-service teachers (e.g., Mertler 2002), where the research instruments used evaluated assessment knowledge, the LAL surveys are in the form of self-report questionnaires. The findings, however, are similar, pointing at deficiencies in teachers' and teacher candidates' assessment knowledge as well as at the lack of proper language assessment training for teacher candidates. The above surveys also draw attention to the often unattainable gap between the idealized and the realized: the declarative broad knowledge base required for performing declarative state-mandated assessment functions contrasted with the teachers' inability to deliver due to limited expertise. In a recent survey Lam (2015), for example, unveils the inadequacy of assessment training in teacher education institutions in Hong Kong, especially in view of local

assessment reforms. A survey conducted among instructors of language testing courses in China showed that contents followed a traditional testing approach with negligible evidence of the assessment culture that the language-testing field was beginning to endorse (Jin 2010). Vogt and Tsagari (2014) found that many aspects of language testing literacy that teachers are expected to possess are underdeveloped and acquired on the job. This is because the majority of the teachers surveyed ($N = 853$ in seven European countries) received little or no training in assessment, and, among other things, also lacked knowledge in self and peer assessment and portfolio use, competencies required for implementing the CEFR framework. Analysis of language teachers' assessment literacy in three national settings yielded differences as to teachers' needs and willingness to partake in assessment training, reinforcing the significance of contextualized considerations including institutional culture when determining LAL contents and modes of acquisition.

Recent research studies reinforce this localized perspective focusing on the language teacher herself, her perceptions, and beliefs, rather than on predetermined language testing content. Csépes (2014) presents research conducted in Hungary reflecting the gap between a recent government policy to move towards "assessment for learning practices" and teachers' reluctance to adopt alternative assessment procedures. Such realizations are leading to growing awareness of the need to consider the interaction between the teachers' beliefs and previous experience and their assessment activities. Scarino (2013) provides accounts of emerging assessment awareness and knowledge amongst language teachers that stem from and incorporate their beliefs, suppositions, and understandings of assessment. Engagement in a reflective process of their own assessment practices facilitates the integration of new knowledge and the creation of a personalized LAL knowledge base. This notion resonates in reports on learner-centered language assessment courses attuned to learners' needs that allow for a broad critical perspective of language assessment and practice (Kleinsasser 2005).

Based on the growing data on assessment needs and practices, a number of definitions of LAL have been offered. Most of the definitions provide general frameworks, some more detailed than others, combining testing and assessment cultures to varying degrees and denoting critical ethical principles (Davies 2008). Mention of tests or the testing process appears in most definitions, while references to other forms of assessment are often absent. The relevance of language issues or competencies is also often precluded from the core knowledge required. For example, the definition by O'Loughlin (2013, p. 363) states that LAL "potentially includes the acquisition of a range of skills related to test production, test score interpretation and use, and test evaluation in conjunction with the development of a critical understanding about the roles and functions of assessment within education and society." Likewise, Pill and Harding (2013, p. 381) see knowledge in language assessment "as indicating a repertoire of competences that enable an individual to understand, evaluate and, in some cases, create language tests and analyze test data." Tests still reign as the overarching assessment tool, perhaps an understandable phenomenon as both the O'Loughlin and the Pill and Hardy research studies focus on test use.

Fulcher provides a more detailed definition that elaborates on the wider assessment framework underlying assessment literate principles and concepts. Similar to the aforementioned definitions, tests, both large-scale and classroom, are highlighted, and no reference is made either to other assessment instruments or to language-assessment features.

> The knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. (Fulcher 2012, p. 125)

Following the definition, however, a breakdown into three categories is offered: contexts, principles, and practices, where the practices are "distinctly language-focused."

Inbar-Lourie (2008b, 2013) acknowledges the primary contribution of general educational assessment literacy to LAL but reaffirms its distinctive language-based concerns. The definition provided therefore takes the form of layers of knowledge, with the bottom layers comprising general assessment literacy that forms a basis for the language-oriented matters. As part of the general assessment knowledge a language assessment literate individual has to know "the reasoning or rationale for assessment (the 'why'), the description of the trait to be assessed (the 'what'), and the assessment process (the 'how')." Nonetheless, the core competences will reflect current views about the social role of assessment in general and language assessment in particular, contemporary views about the nature of language knowledge, and give due emphases to both classroom and external assessment practices. The focus and intensity would vary and depend on the target audience, but an introduction to the core components will be obtained by all participants, including discussion of some of the unresolved controversies and tensions in the field (pp. 396–397). Hence, the language components are underlined, as well as diversity in the scope and depth of the literacy required. Such diversity would depend on the nature of the assessment task and the agents performing it, some of whom may be non-language testing experts.

The need to widen the potential circle of language assessors was expanded and elaborated upon by Taylor (2009, 2013), who draws attention to differential LAL needs. This is apparent, for example, in the case of university admission decisions that administrators make with regards to entrance to higher education (O'Loughlin 2013). Pill and Harding (2013) likewise address this dilemma, in their research on the nature of the LAL parliament members require in order to make sound policy decisions regarding the English language skills of immigrating physicians. The analysis is based on a framework which envisions literacy as a relative and modifiable rather than absolute concept, ranging along a continuum of different levels to the maximum attainment of *multidimensional literacy* (Pill and Harding 2013, p. 383). A follow-up on this idea is elaborated on in Taylor (2013) who outlines

the knowledge elements assessment stakeholders require in different areas. The eight dimensions listed herewith derived from research on LAL form a comprehensive LAL framework: knowledge of theory; technical skills; principles and concepts; language pedagogy; sociocultural values; local practices; personal beliefs/attitudes; and scores and decision-making. In an attempt to operationalize the concept Taylor ties together the LAL dimensions with the literacy continuum offered by Pill and Harding, illustrating via a web of competencies the LAL profile different stakeholders (test writers, classroom teachers, university administrators, and professional language testers), may need. This proposed construct with its intersecting profiles has drawn interest for its potential use in designing the depth and range of assessment initiatives for particular populations.

## Work in Progress

The above discussion of LAL definitions and the recognition of the need to tailor assessment literacy to different audiences carries meaningful ramifications to the field and points at the need for further research, some of which is underway. The report here will focus on three studies: one within the realm of setting an LAL framework for teachers, the second on providing LAL training in response to teacher initiation, and the third on operationalizing the Taylor (2013) model outlined above.

In an attempt to enhance understandings of the complexities and difficulties in attaining LAL among teachers, Xu (2015) presents a tentative situated conceptual framework based on in-depth case study that explored the dynamic and evolving LAL assessment literacy of a language teacher. The approach to the formation of LAL is a contextualized one, which takes into consideration not just the mere training but also the institutional setting using a constructive interpretive epistemology. Xu introduces the term "assessment literacy in practice," whereby the development of teachers' assessment literacy is formed and shaped in an interactive manner alongside other relevant dynamic professional changes which occur in teachers' conceptualization of teaching, learning, awareness of the assessment process, and of oneself as an assessor. The theoretical knowledge base comprises seven components, which includes a merge between assessment matters and pedagogical knowledge. The research calls for a reconceptualization of language assessment literacy and argues "for a sustainable development mindset for language teacher assessment literacy." (Xu 2015). The findings and research direction presented in the framework offer a richer perspective on the intricacies of LAL and its acquisition. For as the researcher suggests, the transfer from attaining LAL in different settings does not automatically occur withstanding the complexity of the classroom and organization as well as personal variables, such as background knowledge and professional conceptions of assessment, reinforcing the motives expressed by Scarino (2013).

Teachers in the LAL and AL research studies are often treated as passive recipients of information which, when delivered, is not necessarily implemented in their classroom assessment practices. A different conceptualization of the teacher's role in acquiring LAL can be found in teacher empowerment studies that

demonstrate teacher activism in reaching out to attain LAL in order to make a difference in the language assessment arena in their own contexts. Brunfaut and Harding (2014) report on an on-going LAL teacher-training project intended at developing English language tests for Luxembourg secondary schools. The project which has lasted thus far over 3 years was initiated by the teachers to bring about change in the national end-of-school leaving English exam which they felt was not aligned with their current teaching. The LAL training enabled the construction of the test, but beyond that it is noted that following the project the assessment literate teachers "now possess skills and knowledge in language assessment which are of an international standard" and "should be viewed as a highly prized group of professionals" in their own context (Brunfaut and Harding 2014, p. 17).

The third interesting research project underway attempts to apply Taylor's (2013) profiling of differential LAL knowledge needs to different stakeholder groups in and outside the language testing community (language teachers, language test developers, language testing researchers, applied linguists, policymakers, and test takers), aiming to develop suitable LAL profiles based on the protagonists' present knowledge, needs, and goals (Kremmel and Harding 2015). The project touches upon the need to identify potential LAL stakeholders not considered part of the assessment circle and pinpoint their LAL needs.

## Problems and Difficulties

Since the conceptualization of LAL is still in its infantile stage it suffers from growing pains, the most notable of which is an identity dilemma. There seems to be a meaningful gap between contemporary theory that upholds assessment culture principles knowledge and skills and its manifestations in the field. The profession is in a state of flux, keeping to the traditional, familiar, and what is perceived as dependable testing knowledge and skills, while at the same time cautiously examining how to combine testing with new notions when disseminating LAL to future experts. This state of mind is particularly conspicuous in view of the bustling developments that constantly offer new paths for matching current thinking in applied linguistics and language pedagogy with novel less institutionalized assessment options. It also creates a gulf between language education policies, which advocate the theory and practice of assessment cultures and the tools and knowledge provided in testing or assessment courses. The dilemma is a profound and goes beyond choice of assessment instruments, for testing and assessment cultures are each anchored in different epistemological paradigms implying differences in LAL contents and orientations. These include different understandings of the role of language in society, whether socially constructed, multidimensional, and dynamic or formulated, monolingual, and static, and the role of the students, whether passive or active in the process of learning and assessment.

Hence, the largest group of LAL consumers, language teachers, as well as their students receive ambiguous messages: asked to conduct formative classroom assessment but unable to deliver, having been trained (if at all) to function in testing-

embedded environments. An example of this tension can be found in the case reported by Inbar-Lourie and Levi (2015) on the use of an integrated teaching and assessment speaking kit intended for junior high EFL learners in Israel. Findings showed that the tool, developed for the ministry of education, was not implemented by the English teachers, though it follows declared formative assessment for learning policies and is aligned with curriculum goals. This was due mostly to the lack of teacher training in formative assessment practices, in using the kit and utilizing the feedback to improve instruction strategies and involve the learners in the process. In addition, ministry officials and decision-makers were not fully aware of the implications of introducing a formatively intended assessment initiative such as the one researched in terms of the knowledge, skills, and principles required.

## Future Directions

The research conducted holds promise as to the future directions of LAL. It has passed the important phase of uncovering the variables, the existing agendas, and needs and is now proceeding towards fine-tuning in terms of operationalizing theoretical frameworks and delving deeper into LAL intricacies in terms of content and research approaches. Just as the previous research in the form of surveys has helped articulate the LAL quandaries, future research can help make meaningful progress as to the nature of the canon needed. However, such research needs to be conducted collaboratively with the stakeholders outside the traditional language testing domain as professionals from different fields of knowledge are involved in language assessment decisions that require situated expertise. Only a shared collaborative effort, one that merges expertise in language assessment with expertise in the local context, can create meaningful assessment solutions to the dynamic issues that arise, especially in view of global changes impacting language use. Thus, there is a need to reach out to new assessors in different capacities not just to disseminate existing knowledge but to create amalgamated understandings as to assessment targets, tools, procedures, analysis and intended but also unintended consequences. Such an interactive negotiated process will enable a reduction to the core basics, the essentials, with added localized knowledge suited to specific needs. Additionally, the assessment circle should be expanded to include consumers of LAL – parents, students, and principals. This will eventually enrich the knowledge base of the testing community and its affiliates at large, aiming for the consideration of localized language assessment literacies rather than promoting a monolithic approach. Exploratory interpretative research approaches need to be part of the research course outlined above, including participatory action research studies that will bring user accounts of LAL practices and needs.

Language teachers, referred to repeatedly as the largest group of LAL stakeholders, are seldom listened to in the LAL debate, though a body of research has surveyed their lacking assessment practices, often the result of inadequate or nonexisting training. There are few research studies which involve teachers' voices from within, looking at the complex variables in the language assessment scene,

educational and assessment policies in and outside the institutions, classroom dynamics, and the language teachers themselves and what they bring to the assessment process. Teachers' research, on its own or collaboratively with assessment experts, needs to be heard loud and clear so as to enrich the literature in knowledge and insight but also to inspire others to take action, try out assessment procedures, and reach conclusions applicable to different situations and individuals. Such insight will contribute to an understanding of the complex LAL puzzle.

To sum up, the envisioned future of LAL is that of a dynamic loose framework, descriptive rather than prescriptive. A framework that sets general guiding principles for different assessment literacies but is aware of local needs and is loose enough to contain them and allow them to develop from theory to practice, but also from practice to theory. A framework where language occupies a central place but where knowledge from other disciplines is welcomed and integrated with joint efforts at creating a truly comprehensive and operationalized assessment entity which reflects rich dynamic and differential language use.

## Cross-References


- ▶ Critical Language Testing
- ▶ Dynamic Assessment
- ▶ Training in Language Assessment
- ▶ Washback, Impact, and Consequences Revisited


## Related Articles in the Encyclopedia of Language and Education


Stephen J. Andrews, Agneta Svalberg: Teacher Language Awareness. In Volume: Language Awareness and Multilingualism

Peter Freebody: Critical Literacy Education: "The Supremely Educational Event". In Volume: Literacies and Language Education

Linda von Hoene: The Professional Development of Foreign Language Instructors in Postsecondary Education. In Volume: Second and Foreign Language Education


## References


American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFTINCMEINEA). (1990). Standards for teacher competence in educational assessment of students. Educational Measurement: Issues and Practice, *9*(4), 30–32.

Barni, M. (2015). In the name of the CEFR: Individuals and standards. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 40–51). New York: Routledge.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa*, 1–12. http://csi.idso.eportalnow.net/uploads/1/1/3/2/11323738/inside_the_black_box_1998.pdf. Accessed 10 June 2015.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 137–143). Cambridge: Cambridge University Press.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing, 25*(3), 349–383.

Brunfaut, T., & Harding, L. (2014). Developing English language tests for Luxembourg secondary schools: The Test Design and Evaluation (TDE) project, 2011–2014. Lancaster University. http://portal.education.lu/Portals/22/English/Documents/BrunfautandHarding2014.pdf. Accessed 10 June 2015.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Council of Europe.

Csépes, I. (2014). Language assessment literacy in English teacher training programmers in Hungary. In J. Horváth & P. Medgyes (Eds.), *Studies in honour of Nikolov Marianne* (pp. 399–411). Pécs: Lingua Franca Csoport.

Davies, A. (1999). The role of the segmental dictionary in professional validation: Constructing a dictionary in language testing. In A. Davies, A. Brown, C. Elder, K. Hill, T. Lumley, & T. MaNamera (Eds.), *Dictionary of language testing* (Studies in language testing, Vol. 7, pp. 242–251). Cambridge: Cambridge University Press.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347.

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393–415.

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education, 34*(5–6), 576–591.

Elder, C. (2015). Acknowledging the diversity of the language learner population Australia: Towards context-sensitive language standards. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 52–64). New York: Routledge.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Basingstoke: Palgrave Macmillan.

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs. Report: Part one – General findings*. http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf. Accessed 6 May 2015.

Heejeong, G. (2013). Defining assessment literacy: Is it different for language testers and nonlanguage testers? *Language Testing, 30*(3), 345–362.

Huhta, A., Hirvalä, T., & Banerjee, J. (2005). *European survey of language testing and assessment needs: Report: Part two – Regional findings.* http://users.jyu.fi/~huhta/ENLTA2/First_page.html. Accessed 6 May 2015.

ILTA Code of Ethics. (2000). http://www.iltaonline.com/index.php/en/resources/ilta-code-of-ethics. Accessed 7 June 2016.

Inbar-Lourie, O. (2008a). Language assessment culture. In E. Shohamy (Ed.), *Language testing and assessment* (Vol. 7). N. Hornberger (General Ed.), *Encyclopedia of language and education* (2nd ed., pp. 285–300). New York: Springer.

Inbar-Lourie, O. (2008b). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402.

Inbar-Lourie, O. (2013). Guest editorial to the special issue on language assessment literacy. *Language Testing, 30*(3), 301–307.

Inbar-Lourie, O., & Levi, Z. (2015, March). *Implementing formative classroom assessment initiatives: What language assessment literacy knowledge is required*? Paper presented at the 37 annual Language Testing Research Colloquium (LTRC), Toronto.
Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555–584.
Kleinsasser, R. C. (2005). Transforming a postgraduate level assessment course: A second language teacher educator's narrative. *Prospect, 20*, 77–102.
Kremmel, B., & Harding, L. (2015, May). *Developing language assessment literacy profiles for different stakeholders – Needs, lacks and wants.* Paper presented at the 12th EALTA conference, Copenhagen.
Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169–197.
Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.
Mertler, C. (2002). *Classroom assessment literacy inventory.* http://pareonline.net/htm/v8n22/cali.htm. Retrieved 30 Aug 2016.
O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380.
Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381–402.
Poehner, M. E. (2009). Dynamic assessment as a dialectic framework for classroom activity: Evidence from second language (L2) learners. *Journal of Cognitive Education and Psychology, 8*, 252–268.
Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265–273.
Purpura, J. E., & Turner, C. E. (2013, March). *Learning-oriented assessment in classrooms: A place where SLA, interaction, and language assessment interface*. ILTA/AAAL joint symposium on "LOA in classrooms", Dallas.
Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–327.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.
Shohamy, E. (2001). *The power of tests*. Harlow: Pearson Education Ltd.
Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal, 95*(3), 418–429.
Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(3), 534–539.
Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36.
Taylor, L. (2013). Communicating the theory, practice, and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.
van Leeuwen, C. (2006). From assessment anecdotes to learning practices. In R. Wilkinson, V. Zegers, & C. van Leeuwen (Eds.), *Bridging the assessment gap in English-medium higher education* (FLF, Vol. 40). Bochum: AKS-Verlag.
Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.
Webb, N. L. (2002, April). *Assessment literacy in a standards-based urban education setting*. Paper presented at the American Educational Research Association Annual Meeting in New Orleans. http://facstaff.wcer.wisc.edu/normw/AERA%202002/Assessment%20literacy%20NLW%20Final%2032602.pdf. Accessed 6 June 2015.
Xu, Y. (2015, March). *Language assessment literacy in practice: A case study of a Chinese university English teacher*. Paper presented at the 37 annual Language Testing Research Colloquium (LTRC), Toronto.

# Language Assessment in Higher Education

Catherine Elder

**Abstract**

This chapter highlights the role of English proficiency in academic study and the associated language assessment issues that emerge in the higher education environment. It considers validity issues surrounding the design and use of the tests used for (i) establishing minimum English entry requirements, (ii) identifying language support needs postentry and/or making English course placement decisions, (iii) establishing readiness to teach academic content through the medium of English, (iv) assessing the adequacy of English proficiency in the context of mainstream academic assignments, and (v) gauging the English standards achieved at exit from the university, including, by implication, students' linguistic readiness to enter the workforce.

It is argued that while research and development initiatives instigated by powerful testing agencies have contributed greatly to our thinking about language and have shaped the field of language assessment as we know it today, many problems remain. There are still uncertainties about how best to define and capture the academic language proficiency construct for testing purposes, in ways which can serve highly diverse student populations in contexts which are increasingly internationalized and technology mediated. Current assessment activities focus too much on English standards at university entry and too little on policies and practices geared to monitoring and fostering students' language development throughout the course of their academic study. Future research and actions which might address some of these challenges are proposed.

C. Elder (✉)
School of Languages and Linguistics, University of Melbourne, Parkville, VIC, Australia
e-mail: caelder@unimelb.edu.au

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_35

**Keywords**
English for academic purposes • Language proficiency • Language testing

## Contents



## Introduction

The shift toward the use of English as a lingua franca in academic and professional contexts worldwide has resulted in a growing emphasis on the role of English proficiency in academic study and on its importance for effective functioning in the subsequent professional workplaces in which university graduates seek employment. Whether English is taught as a subject in its own right, used as a medium of academic instruction, or used as a vehicle for accessing relevant course readings or other research publications, many higher education institutions, both in Anglophone or non-Anglophone countries, stipulate minimum English entry requirements to ensure that admitted students can cope with the language demands of their studies. The need to implement these requirements has opened up commercial opportunities for language testing agencies and at the same time fostered research into the nature of language proficiency and how it is best assessed.

English admissions testing is only part of the story of assessment in higher education however. A range of postentry procedures have been developed in different institutions to identify the language learning needs, whether written or spoken, of admitted students. Language assessment tools may be used for both undergraduate or postgraduate students to exempt them from further English language study requirements or for placement purposes, to determine appropriate class levels or suitable language study pathways in academic writing or learning centers or within mainstream academic programs. Tests may also be used to determine whether graduate students applying to work as international teaching assistants (ITAs) or nonnative English-speaking faculty members working in English-medium institutions have the necessary English competence to manage their role and to diagnose their language support needs.

Although issues of English proficiency loom large at the point of selection and in the early stages of university study, they tend to fade from view as study progresses, due perhaps to under-resourcing and the low status of language and language instructors on the one hand and, on the other, to a lack of understanding or commitment to language issues by institutional leaders and university faculty. Insufficient attention to language matters within university degree programs and confusion about the place of language in the assessment of academic content may adversely affect learning outcomes, hinder language development, and limit future employment opportunities. Some institutions have therefore instituted exit standards of English to address this situation.

This chapter considers the various solutions which have been adopted for pre-entry, postentry, and exit assessments of English proficiency in higher education, the challenges that remain, and the implications of work in this area for the theory and practice of language testing.

## Early Developments

Exploring the English demands of academic study has been the focus of numerous research efforts in applied linguistics since the 1950s when English-medium universities began to open their doors to foreign/international students. The need to ensure that such students were equipped with adequate language skills for academic study triggered the development of a number of influential English language admissions tests designed for large-scale administration. The thinking around those tests has played a big role in shaping the field of language testing as we know it today.

Two histories of language testing, *Measured Words* by Bernard Spolsky (1985) and *Assessing Academic English* by Alan Davies (2008), have charted these early developments in detail with the former focusing primarily on the United States, with particular emphasis on the Test of English as a Foreign Language (TOEFL) developed by the Educational Testing Service (ETS), and the latter on the United Kingdom, with specific reference to the English Language Testing Service (ELTS) and its successor, the International English Language Testing Service (IELTS), developed by the University of Cambridge Local Examinations Certificate (UCLES), which has recently changed its name to Cambridge English Language Assessment. The TOEFL and ELTS/IELTS were products of two very different educational and measurement traditions and the decision-making underpinning their design is characterized by both Davies and Spolsky as a struggle between validity and reliability, with strikingly different outcomes in each case.

The original TOEFL, launched in 1964 by the ETS in Princeton New Jersey, privileged reliability over validity and was more strongly influenced by developments and advances in educational measurement theory than was IELTS. The first form of the test was a 270-item multiple-choice test of 3 h duration covering the skills of structure and vocabulary as well as reading and listening comprehension. The adoption of a discrete-point general proficiency test that was "purely objective,

psychometrically pure, machine-scored and machine-like, cost effective and profitable, secure and efficient" (p. 79) is seen by Spolsky as a lost opportunity and a retreat from more progressive approaches to assessment that were emerging in the United States around that time, including the trend toward direct assessment of writing and/or speaking initiated by the designers of the Michigan Test of English Language Proficiency and the American University Language Centre tests. Although the semi-direct Test of Spoken English (TSE) and the essay-based Test of Written English (TWE) were later developed as optional add-ons to the TOEFL, these initiatives were driven, Spolsky claims, by market demand rather than in response to the tenets of communicative language testing.

The ELTS, by contrast, is described by Davies as representing a bold paradigm shift with respect to the more traditional general proficiency tests which preceded it. The adoption of a communicative approach constituted a radical departure from the Lado-inspired indirect testing of discrete elements of language as exemplified in tests like the English Proficiency Test Battery. The ELTS aimed for authenticity, offering test takers the choice of one of six discipline-specific modules, each with its own listening, reading, writing, and speaking component designed to reflect the language demands of the chosen subject area. Davies argues that the test's emphasis on purposeful and contextualized language use represented a commitment to validity rather than reliability. (Reliability was inevitably compromised by the decision to use a single marker to score the open-ended writing and speaking tasks and by the reported difficulty of developing highly correlated parallel forms of each disciplinary module.) UCLES eventually pulled back from its foray into the until then relatively unexplored terrain of English for Specific Purpose (ESP) testing by first reducing (Alderson and Clapham 1993) and then abandoning its discipline-specific modules. Nevertheless, the new more generalized academic test (IELTS) remained broadly communicative in its various instantiations (Taylor and Falvey 2007).

Regardless of the different orientations of these two pioneer tests of English for academic purposes (EAP), the discussions and research surrounding their development and implementation provided impetus and data for early theory building in the then relatively new discipline of language testing about the nature of the language proficiency construct: whether it should be seen unitary, as Oller (1983) proposed, or multi-componential as Canale and Swain (1980) would have it. Validity studies at this time were mainly statistical in orientation, exploring the factor structure of test items and skill components and the correlations between test takers' scores and various criterion measures of language ability or academic performance in the higher education context.

From the 1990s onward, theoretical models of language proficiency and approaches to test development became more elaborate (see Bachman 1990; Weir 2005). Although the utility of these models for practitioners is sometimes questioned, what is certain is that approaches to developing and validating the two major English admissions tests became more systematic in response to deeper understandings of the language processing and contextual parameters that shape test takers' engagement with and response to test tasks (Taylor 2014). Internally and

externally commissioned research activity burgeoned within both organizations and appeared in research reports (www.ets.org/toefl/research/archives/research_report/), monographs, and edited volumes (www.cambridgeenglish.org/silt/).

## Major Contributions

### Admissions Testing

The concentration of resources within a small number of well-resourced testing agencies means that research relating to large -cale English admissions tests continues to be highly influential in the twenty-first century, featuring prominently in the discipline's premier journals, *Language Testing* and *Language Assessment Quarterly*, and in other applied linguistic outlets.

Two major developments in the first decade of the new millennium are worthy of note. One was the launch of internet-based TOEFL (TOEFL iBT) in the late 2005 after years of preparatory work including a series of conceptual framework documents and empirical studies (see www.toefl.org) specifying the constructs of academic speaking, listening, reading, and writing proficiency in communicative terms. The research surrounding this test, most notably a monograph by Chapelle et al. (2008), has firmly put to rest any notion that validity is not a primary concern.

The second development was the 2009 launch of the computer-based Pearson Test of English (PTE) (academic), now recognized by many English-medium institutions as an alternative to IELTS and TOEFL iBT. The PTE is designed to measure English competence in academic context based on reading, listening, speaking, and writing tasks but, unlike its competitor tests, it also measures the "enabling skills" of vocabulary, grammar, spelling, pronunciation, and discourse competence, offering information about these elements in a separate report. Validation research on this newcomer test is starting to appear (see http://pearsonpte.com/research/), although few studies on this test have thus far been published in the major peer-reviewed journals.

Both PTE and the TOEFL iBT have followed in the footsteps of the earlier ELTS test in opting for integrated tasks, intended to reflect the integration of different language skills in the academic environment. However, these are operationalized very differently for each test and without any discipline-specific branching of the kind previously attempted by the ELTS test designers. This trend toward integrated tasks is generating a strong body of validation research (e.g., Cumming et al. 2006; Wei 2012; Kyle et al. 2015). In addition, both the ETS and Pearson have developed proprietary automated scoring systems for rating constructed speaking and written responses. This constitutes a major innovation in the language testing arena and is stimulating considerable research interest (Xi 2010).

A range of other English tests, both large and small scale and too numerous to name here, is also operating in the admissions testing marketplace, whether designed for local or international use. Among the better known of these are the College English Test (CET) in mainland China (Zheng and Cheng 2008), the General

English Proficiency Test (GEPT) (Roever and Pan 2008) in Taiwan, and the EIKEN test in Japan (Dunlea 2010). With the proliferation of English language admissions tests has come the need for equivalence studies which align the tests with one another or with a scale, such as the Common European Framework of Reference (CEFR) now used by many institutions both within and outside Europe for policy-making, goal setting, and reporting purposes (e.g., Pearson 2010; Educational Testing Service 2010). While claims of equivalence between tests with very different qualities are at best approximations, making these links is a political and bureaucratic necessity, given that many jurisdictions accept scores from multiple sources as meeting their entrance requirements and failure to specify such links may result in particular tests being excluded from consideration. The tension between professionalism and pragmatism is particularly acute when commercial interests are at stake.

A heightened emphasis on score utilization and test consequences influenced by the work of Messick (1998), Kane (2012), and Bachman and Palmer (2010) is evident in a growing body of research on washback from high stake admissions tests (e.g., Green 2007; Wall and Horak 2011) and on score users' perceptions. Examples of the latter are studies by O'Loughlin (2011) and Ginther and Elder (2014) of score users' understandings and interpretations of TOEFL iBT, IELTS, and PTE in Australian and American universities. Both studies reveal confusion among academics regarding the meaning of minimum cut scores set by their institutions and a tendency to blame the admissions tests for what are perceived to be unduly low English standards among their students. Such findings highlight the importance of effective institutional policy-making surrounding admissions test score use. They also signal the need to build language assessment literacy among score users, an area that is attracting increasing research attention (e.g., see Inbar 2013).

## Alternative Routes for University Admission and Postentry English Language Assessment

While much attention has been paid to the design delivery and validation of English admissions tests and to the importance of appropriate uses of test scores, there remain large numbers of international and other students who enter the university via pathways which do not require satisfactory performance on a high-stakes language test. These students may participate in foundational or EAP courses geared to preparing them for academic study, where satisfactory completion is recognized as adequate grounds for admission (or provisional admission in some cases). While there are strong pedagogical arguments for preparatory courses, there is also a risk that administrators may set entry levels too low with a view to attracting fee-paying clientele. Prevailing questions for such courses are whether the content and quality of their in-house formative and summative assessments, which are often designed by staff without professional expertise in language assessment, are well-suited to determining readiness for academic entry and whether students perform academically on a par with those entering via the usual admissions testing pathway (e.g., Owen 2012; Cross and O'Loughlin 2013; Heitner et al. 2014).

Regardless of the pathway taken, many students meeting minimum entrance requirements may face language challenges in higher education. This applies not only to international students but also to domestic ones, particularly those from low English literacy backgrounds or who have specialized in school subjects that make limited language demands. Postentry procedures are therefore needed to determine who may need assistance and what form this assistance should take. Read (2015) devotes an entire volume to describing design and implementation issues surrounding postentry English language assessments (PELAs), administered to students following admission to the university, with the aim of identifying those who may be linguistically at risk, regardless of their language background. Many such tests are locally developed and function broadly as placement tools to allocate learners to appropriate English language development programs. Other institutions catering largely for second language learners make use of generic tools such as Accuplacer Companion (Johnson and Riazi 2015) or the TOEFL (Kokhan and Lin 2014) for this purpose. Some PELAs claim to be diagnostic and offer learners the opportunity to monitor their progress over time (e.g., Urmston et al. 2013). An interesting strand of recent research, which potentially obviates the need for custom-built postentry assessments, looks at how tests designed initially for admissions or placement purposes can be retrofitted to provide more fine-grained diagnostic feedback to learners (Fox 2009; Jang 2009; Li 2011; Kim 2015).

Whether these diverse approaches to PELA actually achieve their intended purpose of effectively identifying and addressing language needs requires ongoing validation efforts that have been slow to emerge in the PELA environment. An argument-based validation framework for PELAs devised by Knoch and Elder (2013) positions the testing instrument as only one component of a larger institutional policy and program dedicated to the provision of appropriate interventions for at-risk students and to the ongoing monitoring of these interventions.

## Assessing English Proficiency for University Teaching

In addition to assessments which target the academic English needs of students for study purposes are those designed to determine whether graduate students or indeed other nonnative English-speaking staff members employed in English-medium institutions have adequate oral proficiency for teaching purposes. In the United States, it has long been the practice to employ international graduate students as teaching assistants (ITAs), particularly in undergraduate courses in science, mathematics, and engineering. A variety of solutions have been adopted for screening and subsequent training of ITAs including the use of existing scores on the speaking component of an international admissions test like the TOEFL iBT on the one hand (see validation study by Xi 2008) and, on the other, custom-built tests developed for local contexts such as the Test of Oral Proficiency (TOP) at the University of California, Los Angeles (Farnsworth 2013). The chief benefit of locally developed procedures is that they can target with more precision the communication skills of particular relevance for teaching. Test results can also

be better integrated with local instructional programs for those requiring further training.

In Northern Europe, where increasing numbers of English-medium degree programs are offered to graduate students, teacher language proficiency assessments may also be required. One assessment tool used in this context is the Test of Oral English Proficiency for Academic Staff (TOEPAS), an EAP certification test developed at the University of Copenhagen that also offers formative feedback on classroom performance (Dimova and Kling 2015). Even where such feedback is provided, the authors point out, a test on its own is unlikely to bring about a change in practice in the absence of appropriate policies incentivizing English language development. Indeed, there may be resistance to such testing among academic staff who see such initiatives as imposing native-speaker norms in an English as a lingua franca environment where such standards are deemed inappropriate and unnecessary (see further discussion below).

## The Role of English in Academic Outcomes

Efforts made to measure the English proficiency of staff and students, whether prior to or following admission to university, are predicated on the notion that insufficient English will hamper progress and adversely affect study outcomes. Research into the predictive power of language tests has however produced mixed findings (e.g., see Graham 1987; Vinke and Jochems 1993; Allwright and Banerjee 1997; Dooey and Oliver 2002), showing that language proficiency seldom accounts for more than 10% of variance in academic performance (however measured) and that its role may vary according to discipline. Many reasons have been proffered to explain this limited predictive power of language: the fact that language proficiency is only one of a host of factors contributing to study outcomes, that the contexts of investigations and the role of proficiency within these contexts varies as does the size of the student sample investigated, that the proficiency range of the sample is truncated because only admitted students are included, that the criteria for measuring success are crude and unreliable, and that the correlational measures normally used for such studies are difficult to interpret (Cho and Bridgeman 2012).

What also hinders efforts to explore the role of language proficiency in academic success is a lack of transparency in the mechanisms for assessing academic achievement in the higher education context (Knight 2002). The diversity of the university student population poses particular challenges for assessment. O'Hagan (2014), for example, provides evidence of bias in faculty judgments of essays produced by students from English- and non-English-speaking backgrounds and a disturbing level of inconsistency in marks assigned. Weigle (2002) observes a degree of uncertainty among university assessors about the extent to which English language issues should figure in evaluations of essay quality.

Such findings raise concerns about the equity and fairness of assessment in the increasingly internationalized and technology-mediated higher education context

and suggest that students may be receiving mixed messages about the role of language in academic performance.

## Assessing Language Skills at Exit from the University

Uncertainty among academics about how or indeed whether to attend to language in assessing disciplinary content may result in students failing to pay due attention to their English development. While the idea of language intervention to support at-risk students has been embraced in many English-medium universities, such interventions tend to be confined to the early stages of undergraduate study and not necessarily embedded within the core academic curriculum. Moreover, feedback on course assignments varies widely and seldom makes mention of language issues (O'Hagan 2014). English language development over the duration of an academic degree program cannot therefore be taken for granted, as a number of recent pre- and posttest studies spanning different study periods have shown (e.g., Knoch et al. 2015; O'Loughlin and Arkoudis 2009). Of course it must be borne in mind that a number of factors independent of the academic context, such as attitudes to English and English speakers and exposure to English outside the classroom, may hinder language development. On the other hand, it is also conceivable that the general proficiency tests often used for comparison purposes may be insensitive to students' growing mastery of discipline-specific language genres.

Be that as it may, some universities have seen fit to introduce exit tests of English language proficiency to motivate university students to improve their English and to provide future employers with information about students' levels of proficiency at the time of graduation. One such test, expressly designed for this purpose, is the Graduating Students Language Proficiency Test (GSLPA), a task-based assessment designed to mirror the demands of Hong Kong work situation (Qian 2007) and to generate positive washback on teaching and learning in the university context.

More commonly, however, a one-size-fits-all test is used to gauge exit standards, such as the General English Proficiency Test (Roever and Pan 2008) in Taiwan. Likewise in those English-speaking countries where exit standards are formally monitored more often than not it is high currency English admissions tests like IELTS or TOEFL that are chosen, rather than measures targeting the skills required for workplace communication. Such tests are not linked to any teaching syllabus (unlike the College English Test (Zheng and Cheng 2008) in China, for example, where the goals of language learning and expected levels of proficiency at different stages of academic study are made explicit). The impact of these high-stakes tests on the teaching and learning of English in higher education contexts remains uncertain as does the utility of the information they provide for employers.

The recently launched Global Scale of English (GSE) (http://www.english.com/gse#.VsuZExEz7ww) may prove useful in this regard. The GSE is a standardized granular scale (mapped onto the broader CEFR levels) which profiles the English language learning trajectory of learners in a series of small steps, each linked to a

precise set of teaching and learning objectives. The Pearson group has developed versions of this scale for both academic and professional contexts with a view to measuring and profiling language learning progress in a manner that is meaningful to users.

## Work in Progress

While research into language use in academic settings has informed the design of all the major English admissions tests, it remains a priority area for research, suggesting that the quest for authenticity, considered part and parcel of communicative language testing, is not easily satisfied. Recently funded projects address topics such as strategy use in IELTS reading test tasks compared with those used in academic study and the comparability of students' performance on the various components of TOEFL iBT with university reading, writing, and speaking requirements. Findings of such studies serve to explore the construct and content validity of current tasks as well as to identify areas for test enhancement.

A contentious issue, also connected to the notion of test authenticity, is whether the norms governing English assessment either before, during, or on completion of an academic degree are reflective of the current communicative realities in increasingly culturally and linguistically heterogeneous academic contexts where the majority of students and many staff speak English as an additional language. Work on English as a Lingua Franca in academic settings (ELFA) (Mauranen 2013) shows that ELFA interactions tend to be managed in unconventional ways which do not conform with the native-speaker norms that underlie traditional English language assessment. Canagarajah (2006) proposes that it is the skills of adaptability, which language users need to cope with different varieties of English, that should be the focus of assessment rather than mastery of any single variety. While the implications of these changes are yet to manifest in operational assessment tools, some current research, funded under the TOEFL grants program (https://www.ets.org/toefl/grants/recipients#coe) offers insights into how the measurement of the intelligibility of different varieties of English might inform the design of listening assessments. Newbold (2015) considers the possible role of learner and ELF corpora in the identification of a test construct for oral production. Harding (2015), in a promising development, lays out a blueprint for the design of purpose-built ELF assessment tasks, which will be operationalized with examples relevant to the academic environment.

A further area of current enquiry concerns the alignment between tasks used for university assessment purposes and language use in professional contexts. One example is a study by Knoch et al. (2016) which finds considerable dissonance between the writing demands of the final year of university study and what is expected of graduating engineering and accounting students in the first year of their professional working lives. Findings from studies like these have implications both for the design of university assessment tasks and also for the tests that are chosen to measure exit standards in higher education. New initiatives such as

the ACT21S project (Griffin and Care 2015), which attempts to specify in measurable terms the key understandings and skills needed by productive and creative workers and citizens of the twenty-first century, may assist language assessors in devising innovative methods of assessment more attuned to future workplace demands.

## Problems and Difficulties

The above overview has uncovered a number of unresolved problems relating to English assessment in higher education. Perhaps the most central of these is that of construct representation. How is the complex and multilayered construct of English proficiency for academic purposes with all its sociolinguistic and disciplinary variation best represented in a language test? And to what extent should the construct of academic English be expanded to encompass the language demands of the future professional settings which students are being prepared for? The shifting formats and changing content of the various English admissions tests reviewed above reflect serious attempts to resolve these questions, but convincing evidence that one approach has greater predictive power than another is lacking and difficult to gather given the plethora of other variables involved in academic success.

The many alternative pathways to university entry with their different means of determining linguistic readiness for academic study create further complexity. Postentry English assessments have attempted to flag linguistically at-risk students who might otherwise slip through the net unnoticed, but such tests run the same risks of construct under-representation as the admissions tests described above, along with the burden of validity evidence required to support their claim to identify the language learning needs of a diverse student population with highly variable proficiency profiles.

The extent to which students develop their English language resources during their time at university is a matter of growing concern. While many universities acknowledge the importance of providing language enrichment opportunities for their students, particularly at the early stage of academic study, there is limited formal monitoring of the effectiveness of the interventions provided. Moreover, the place of language in assessing achievement within mainstream academic courses is seldom made explicit either in assessment rubrics or in feedback given to students. There appears to be a lack of shared knowledge of academic standards as well the assessment expertise needed to implement valid and consistent evaluation regimes. Thus, the determination of academic outcomes risks being either a random or biased process, whereby standards vary wildly and students with limited English may be unfairly penalized without due acknowledgement of the cause.

Any failure to adequately address the English language needs of the student population during the course of the academic study cycle in turn creates pressure when students transition from the university to the workplace where limited English may put them at a serious advantage. Such pressure is compounded by the fact that the tasks used in course assignments and the tests used to gauge exit standards are not always aligned with employers' expectations.

## Future Directions

It would seem that the key to more valid and useful English assessment in higher education lies in effective long-term institutional policy-making which puts language tests in their rightful place as part of an integrated teaching and learning program which makes explicit not only the content objectives of courses and the standards of achievement expected but also the nature and level of English proficiency expected at various points in students' study trajectory. Effective policy-making, if concerns for language development are to be anything more than tokenistic, must stipulate not only the mechanism for ensuring adequate proficiency at entry but also the means of identifying learning needs and monitoring language development throughout students' courses and on graduation. Such policies should provide the basis for informed choices of assessment tools and intervention strategies as well as outlining strategies for education of score users (including administrators and academics) in interpreting and acting appropriately on assessment information.

As for research, it seems that in a world where the nature and modes of language use are constantly changing, work must continue on construct definition and on the design of innovative task types and assessment criteria that capture the complex amalgam of expressive resources, technological know-how, and disciplinary knowledge and strategies required for effective communication not only in the academic environment but also in the professional domains that students will enter on completion of their studies. A key challenge for designing language assessments in higher education is getting the balance right between judgments of the adequacy of language as vehicle for getting the message across and evaluations of the quality of the message itself. How these different elements combine and the weightings accorded to each will depend on the purpose of the assessment and who stands to benefit from the information the assessment yields – whether this is the institution making selection or placement decisions, the learner or teacher planning strategies for language enrichment, the faculty members assessing achievement of course objectives, or the future employer choosing those who are best equipped for the communicative demands of their professional role. Different design solutions and validity evidence are required for each of these purposes to maximize measurement precision and beneficial washback effects. Collaboration between language and content experts in these test development and validation efforts is also essential.

Finally, given the linguistic and cultural diversity characteristic of most higher education environments, and the fact "no-one is a native speaker of the specialist domain of academic English " (Mauranen 2013, p. 13), new assessment tools may need to rely on norms defined not by native speakers but by competent ELFA users in the context of concern. Whether such norms are generalizable across different ELFA contexts is a matter for further exploration.

## Cross-References

- ▶ Assessing Students' Content Knowledge and Language Proficiency
- ▶ Cognitive Aspects of Language Assessment
- ▶ The Common European Framework of Reference (CEFR)
- ▶ Training in Language Assessment

## Related Articles in the Encyclopedia of Language and Education

Patricia Duff: Language Socialization, Higher Education and Work. In Volume: Language Socialization

Linda von Hoene: The Professional Development of Foreign Language Instructors in Postsecondary Education. In Volume: Second and Foreign Language Education

Mary R. Lea: Academic Literacies in Theory and Practice. In Volume: Literacies and Language Education

Fredricka L. Stoller, Shannon Fitzsimmons-Doolan: Content-Based Instruction. In Volume: Second and Foreign Language Education

## References

Alderson, J. C., & Clapham, C. (eds). (1993). Examining the ELTS Test: An account of the first stage of the ELTS Revision Project – Research Report 2.

Allwright, D., & Banerjee, J. (1997). *Investigating the accuracy of admissions criteria: A case study of a British university*. Lancaster: Center for Research in Language Education, Lancaster University.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3*(3), 229–242.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 29*(3), 421–442.

Cross, R., & O'Loughlin, K. (2013). Continuous assessment frameworks within university English Pathway Programs: Realizing formative assessment within high stakes contexts. *Studies in Higher Education, 38*(4), 584–594.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdozy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype*

*written tasks for the new TOEFL* (TOEFL Monograph Series, Vol. MS-30). Princeton: Educational Testing Service.
Davies, A. (2008). *Assessing academic English: Testing English proficiency 1950–89*. Cambridge: Cambridge University Press.
Dimova, S., & Kling, J. (2015). Lecturers' English proficiency and university language polices for quality assurance. In R. Wilkinson & M. L. Walsh (Eds.), *Integrating content and language in higher education: From theory to practice selected papers from the 2013 ICLHE conference* (pp. 50–65). Frankfurt: Peter Lang International Academic Publishers.
Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of IELTS as an indicator of future academic success. *Prospect, 17*(1), 36–64.
Dunlea, J. (2010). The Elken Can-do list: Improving feedback for an English proficiency test in Japan. In L. Taylor & C. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (Studies in Language Testing, Vol. 31). Cambridge: Cambridge University Press.
Educational Testing Service. (2010). *Linking TOEFL iBT scores to IELTS scores: A research report.* https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf. Accessed 26 June 2015.
Farnsworth, T. (2013). Assessing the Oral English abilities of international teaching assistants in the USA. In A. Kunnan (Ed.), *Companion to language assessment*. Chichester: Wiley.
Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes, 8*, 26–42.
Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of TOEFL iBT, IELTS (Academic) and the Pearson Test of English in U.S. and Australia: A case study of two university contexts*. Final Report (CEO-RFP 2009–33).
Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21*(2), 505–521.
Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policies and Practice, 14*(1), 75–97.
Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills. Methods and Approach* (pp. 3–33). Dordrecht: Springer.
Harding, L. (2015). Adaptability and ELF communication: The next steps for communicative language testing? In J. Mader & Z. Urkun (Eds.), *Language testing: Current trends and future needs*. IATEFL TEASIG.
Heitner, R. M., Hoekje, B. J., & Braciszewski, P. L. (2014). Tracking English language proficiency and IELTS test scores in an international undergraduate conditional admission program in the United States. In C. J. Linton & L. V. Amoroso (Eds.), *Measured language: Quantitative studies of acquisition, assessment and variation*. Washington, DC: Georgetown University Press.
Inbar, O. (2013). Guest editorial to the special issue on language assessment literacy. *Language Testing, 30*(3), 301–307.
Jang, E. (2009). Cognitive diagnostic assessment of L2 reading ability. Validity arguments for Fusion model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73.
Johnson, R. C., & Riazi, A. M. (2015). Accuplacer Companion in a foreign language context: An argument-based validation of both test score meaning and impact. *Papers in Language Testing and Assessment, 4*(1), 31–58.
Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.
Kim, A.-H. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test. Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258.
Knight, P. (2002). Achilles heels of quality: The assessment of student learning. *Quality in Higher Education, 8*(1), 107–115.
Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessment. *Papers in Language Testing and Assessment, 2*(2), 1–19.

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing, 28*, 39–52.

Knoch, U., May, L., Macqueen, S., Pill, J., & Storch, N. (2016). *Transitioning from university to the workplace: Stakeholder perceptions of academic and professional writing demands*. IELTS Research Report.

Kokhan, K., & Lin, C.-K. (2014). Test of English as a Foreign Language (TOEFL): Interpretation of multiple score reports for ESL Placement. *Papers in Language Testing and Assessment, 3*(1), 1–23.

Kyle, C., Crossley, S., & McNamara, D. (2015). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*. doi:10.1177/0265532215587391.

Li, H. (2011). A cognitive diagnostic analysis for the MELAB reading test. *Spaan Fellow Working papers in Second or Foreign language Assessment, 9*, 17–46.

Mauranen, A. (2013). Exploring ELF: Academic English shaped by non-native speakers. *TESOL Quarterly, 47*(2), 431–433.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1), 35–44.

Newbold. (2015). http://virgo.unive.it/ecf-workflow/upload_pdf/ELLE_4_1_2015_002_Newbold.pdf. Accessed 6 Aug 2015.

O'Hagan, S. (2014). *Variability in assessor responses to undergraduate essays: An issue of assessment quality in higher education*. Bern: Peter Lang.

O'Loughlin, K. (2011). The interpretation and use of test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146–160.

O'Loughlin, K., & Arkoudis, S. (2009). Investigating IELTS exit score gains in higher education. *IELTS Reports, 10*, 95–180.

Oller, J. (1983). *Issues in language testing research*. Rowley: Newbury House Publisher.

Owen, N. (2012). *Can PTE be used as an exit test for a course of academic English?* http://pearsonpte.com/wp-content/uploads/2014/07/Owen_Executive_Summary.pdf

Pearson. (2010). *Aligning PTE Academic test scores to the Common European Framework of Reference for languages.* Research Note. http://pearsonpte.com/research/Documents/Aligning_PTEA_Scores_CEF.pdf. Accessed 12 Aug 2015.

Qian, D. (2007). Assessing university students: Searching for an English language exit test. *RELC Journal, 36*(1), 18–37.

Read, J. (2015). *Assessing English proficiency for university study.* Basingstoke: Palgrave Macmillan.

Roever, C., & Pan, Y. C. (2008). Test review: GEPT; General English Proficiency Test. *Language Testing, 25*(3), 403–407.

Spolsky, B. (1985). *Measured words: the development of objective language testing*. Oxford: Oxford University Press.

Taylor, L. (2014). General Language Proficiency (GLP): Reflections on the "Issues Revisited" from the perspective of a UK Examination Board. *Language Assessment Quarterly, 11*(2), 135–151.

Taylor, L., & Falvey, P. (Eds.). (2007). *IELTS Collected Papers: Research in speaking and writing assessment* (Studies in Language Testing, Vol. 19). Cambridge: Cambridge University Press.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency. *Hong Kong Journal of Applied Linguistics, 14*(2), 60–82.

Vinke, A. A., & Jochems, W. M. G. (1993). English proficiency and success in international postgraduate education. *Higher Education, 26*, 275–285.

Wall, D., & Horak, T. (2011). *The TOEFL Impact Study: Phase 3, the role of the coursebook, and phase 4, describing change.* Final report on long-term study of changes in the TOEFL on teaching in a sample of countries in Central and Eastern Europe. TOEFL iBT Research Series No. 17. http://www.ets.org/research/policy_research_reports/publications/report/2011/jaqc

Wei, W. (2012). Can integrated skills tasks change students' use of learning strategies and materials? A case study using PTE academic integrated skills items. Retrieved 14 Aug 2015 from pearsonpte.com/wp-content/uploads/2014/07/Summary_WeiWei.pdf

Weigle, S. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs*. Princeton: Educational Testing Service.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291–300.

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing, 25*(3), 408–417.

# Language Assessment in Indigenous Contexts in Australia and Canada

Beverly Baker and Gillian Wigglesworth

**Abstract**

Here we present a brief portrait of language assessment in Indigenous contexts in Australia and Canada – two countries who, despite their distance, share important similarities in terms of their historical mistreatment of Indigenous peoples and their languages. Fortunately, both countries are currently experiencing renewed and increasing interest in Indigenous language revitalization. An examination of public school curricular documents in Canada reveals increased attention to development of more Indigenously-informed and innovative assessment practices. In Australia, while there are community-based programs specifically tailored to language revitalization, formal language teaching contexts programs are still geared towards students meeting specific linguistic criteria. However, both contexts reveal a growing acknowledgment of the importance of community involvement in assessment. An examination of both contexts also reveals the challenge of effectively communicating to non-Indigenous educators, policy makers, and other audiences about Indigenous approaches to language assessment.

**Keywords**

Indigenous language assessment • Aboriginal language revitalization • Alternative approaches to Indigenous language evaluation

B. Baker (✉)
Official Languages and Bilingualism Institute, University of Ottawa, Ottawa, ON, Canada
e-mail: Beverly.Baker@uottawa.ca

G. Wigglesworth
Research Unit for Indigenous Language, ARC Centre of Excellence for the Dynamics of Language, Faculty of Arts, University of Melbourne, Parkville, VIC, Australia
e-mail: g.wigglesworth@unimelb.edu.au

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_20

## Contents



## Introduction

Despite the distance between them, Canada and Australia share a great deal in terms of their colonial histories and the resulting treatment of their Indigenous minorities. Both are widely dispersed in terms of geography and both were home to what was historically a vast diversity of Indigenous peoples who spoke a large variety of languages. Both countries have a history of active repression of these languages in favor of the colonial languages (English, and also French in the case of Canada). The result has been that many of the Indigenous language varieties have disappeared, and the remainder are endangered. For example, in Australia, of the original 250–300 languages, only around 100 continue to be spoken, with only 20 still being learned by children as a first language. In Canada, from hundreds of original languages there are between 50 and 90 currently being spoken, depending on the source being consulted; all are classified as endangered by UNESCO (2011; also see Patrick 2010)[1]. However, with the increasing recognition of the importance of the revitalization of Indigenous languages both from a linguistics point of view and for the well-being of these populations, both of these countries are experiencing renewed and increasing interest in Indigenous[2] language revitalization.

Language assessment has received much less attention than pedagogy in these initiatives, but an examination of public school curricular documents in Canada reveals increased attention to development of more Indigenously-informed and innovative assessment practices. In Australia, while there are community-based programs specifically tailored to language revitalization, formal language teaching contexts programs are still geared towards students meeting specific linguistic

[1]The state of revitalization efforts of Canada's indigenous languages is covered in greater detail elsewhere in this Encyclopedia (Also see Onowa McIvor and Teresa L. McCarty).

[2]In this chapter, the terms Indigenous and Aboriginal are used interchangeably to refer to the First Nations, Métis and Inuit peoples in Canada. In Australia, following Hudson et al. (2010), the term Indigenous refers to Aboriginal or Torres Strait Islander people; Aboriginal refers to Australian Aboriginal people.

criteria. However, both contexts reveal a growing acknowledgment of the importance of community involvement in assessment.

## The Case of Canada

### Background

According to data collected by Statistics Canada in 2011, only about 17% of people who report an aboriginal identity also report that they can conduct a conversation in an aboriginal language (Langlois and Turner 2013). This is hardly surprising: from the early 1900s, an explicit government policy of assimilation led to tens of thousands of children being removed from their communities and placed in residential schools, where their languages and cultural practices were suppressed. A recent Canadian government-sponsored Truth and Reconciliation Commission (TRC) has documented thousands of reports of this cultural suppression in residential schools, in addition to other abuses (TRC 2015). Among the report's 94 recommendations were several related specifically to language rights, including a call to introduce an Aboriginal Languages Act and a call to create more postsecondary programs in Aboriginal Languages (TRC 2015). These recommendations add urgency and attention to an already growing movement of aboriginal language revitalization in Canada.

Aboriginal language education in Canada is currently delivered through the following channels:

- Language and culture programs operated through either public schools (where numbers warrant) or band-operated (on-reserve) schools
- Bilingual models, primarily in the north, characterized by home language schooling in early years followed by a gradual increase in French or English instruction
- Early childhood immersion programs following a language nest model, where "fluent language speakers (teachers and Elders). . .speak only the First Nations language to participating children" (Wilson 2004)
- University-level second language courses
- Informal grassroots revitalization initiatives, such as summer language camps, weekend camps, and courses in "Friendship Centers" in urban areas

Aboriginal language protection in Canada is subject to challenges regarding jurisdiction. Education is decentralized in Canada, with each province and territory responsible for developing its own curriculum. However, education on reserves is also supported through the Federal Ministry of Aboriginal Affairs and Northern Development (formerly Indian Affairs). While Canadian provinces and territories have some documents related to Indigenous language curriculum for K-12 public education, some provinces have ceded jurisdiction for educational content to reserves.

## Indigenous Language Assessment as Revealed Through Curriculum Documents

An examination of some provincial and band-created curriculum documents for public school Indigenous language education reveals some explicit comments on Indigenous approaches to assessment. Other documents do not reference assessment directly, instead referring indirectly to the elements of language that are most valued. Document analysis is limited in providing insight into assessment beliefs and practices at the tertiary level, however. An examination of ten university syllabi from across the country reveals that very little explicit detail is provided on evaluation practices. At universities, Indigenous language courses are generally offered either by departments of linguistics (where course content is focused more on language analysis) or in teacher education or Aboriginal Studies programs, where courses are often more conversation-based and can be partially held outside the classroom on the land. Regardless of the focus of teaching and learning, most courses included require major written assignments or examinations and are graded with primarily linguistic criteria.

While all Canadian provinces and territories have an Indigenous language curriculum (See, e.g., The Government of Manitoba 2007; The Government of Ontario 2001; The Government of the Northwest Territories 2015), some are more elaborate in their discussion of assessment. The Kwayaciiwin Education Resource Centre (2014a, b), serving over 20 communities in the province of Ontario, provides ideas on classroom assessment borrowed from the provincial Ministry of Education and also includes an explicit statement on an Indigenous approach to assessment based on Apprenticeship pedagogy (See also Government of Ontario 2001).

Other provinces and territories deal with assessment issues indirectly. For example, the Government of the Northwest Territories (2015), which supports school language programs in Dene and Inuit language, contain detailed learning outcomes by grade level which can be used for assessment purposes. These learning objectives are linguistic (e.g., "identify word patterns"), functional (e.g., "engage in storytelling"), and behavioral (e.g., "be humble"). In addition, the Western Canadian Protocol for Collaboration in Basic Education (2000), which was created with the collaboration of several western provinces, offers a common curriculum framework for K-12 and outlines language use as related to "kinship (respect in relationships), protocol (conduct in ceremonies and social interaction), medicine (personal habits and practice in relation to health and spiritual gifts), ceremonies (roles and conduct), copyright (earning the right to knowledge) and oral tradition (expression of knowledge, its forms and ownership)" (p. 15).

There are commonalities found among all these curriculum documents in the suggested content of assessment as well as the ideal processes for assessment. In terms of content, value is placed on culturally relevant language functions, such as using the language to tell stories, to spiritualize, to express emotion, and to joke. In fact, the use of language for humorous purposes, such joking and teasing, is much more explicitly valued than one sees in non-Aboriginal Canadian language curricula (see Spielmann 1998, and Fagan 2001 for discussions of the importance of humor in

the Canadian Aboriginal context). In terms of the processes of assessment, all the curriculum documents that were examined emphasize the creation of a positive and supportive environment for learning, with avoidance of negative feedback. For example, the Kwayaciiwin documents (2014a) state that "[i]n traditional aboriginal contexts, there are never any negative communications about slow learners, only positive acknowledgments of the process each individual makes" (p. 19). Alberta Education provincial documents (2009), in their discussion of the application of Cree values to education, state that "[t]his is learning where positive feedback, not negative, is given" (pp. 4–5), and advise teachers to "[k]eep the evaluation of work gentle and encouraging" (p. 37). Northwest Territories Inuit language curriculum documents (1996) state that in traditional Inuit approaches to evaluation, feedback from adults is always positive.

Another common theme in the Canadian documents – which is less evident in Australia – is the influence of Master-Apprentice pedagogy on assessment practices. In Master-Apprentice pedagogy, learners work with teachers to determine individualized learning paths – which include the nature and timing of evaluation. As stated in the Alberta Education documents (2009), "The traditional Cree teaching and learning model emphasizes mastery before evaluation, and within that process the student is set up for success" (p. 36). The Kwayaciiwin documents (2014a) emphasize that "[a]ssessment must reflect Indigenous priorities on doing things, inferring what is known about the process from these active demonstrations. . .. It is common to work on a skill at the entry level until one or many Elders acknowledge you and say you are ready to move up" (p. 19).

## The Case of Australia

### Background

Of the original 250–300 languages spoken in Australia (and almost double the number of dialects), only around 11% of the Indigenous population report speaking their traditional language (ABS 2011; McKay 2011). Many of these languages are only spoken by older people and will cease to be spoken once they pass on; thus, the imperative for documentation of these languages is critical and has been a major focus in Australia. Equally critical is the fact that only about 20 traditional Indigenous languages continue to be learned by children as a first language, which are typified by relatively small populations of speakers (Marmion et al. 2014).

As in Canada and the USA, explicit government policies in Australia from the 1890s to the 1970s removed many Aboriginal and Torres Strait Islander children from their families and situated them in missions, residential schools, or foster families where their traditional languages and cultures were either explicitly banned or at best were not nurtured. This contributed significantly to the processes of language loss.

Indigenous languages in Australia are increasingly taught through various types of programs, all of which are also found in Canada:

- Bilingual program in schools in Indigenous communities where the traditional language is the first language of the children who attend the school
- Second language programs in state schools at both primary (elementary) and secondary level in some states, as well as some language revitalization programs
- TAFE (Technical and Further Education) institutions, and university-level courses teaching Indigenous languages
- Community-based language revitalization and reclamation activities, where the language is highly endangered with very few, if any, remaining speakers

## Indigenous Language Assessment as Revealed Through Curriculum Documents

Indigenous languages are taught in some, but by no means all, universities in Australia, generally for credit. As in Canada, on the whole, little information is provided for the assessment of subjects beyond the broad outline of online quizzes, written assignments, oral assessments, and translations, which are designed to assess language proficiency (as with other languages), but which also usually include a component of cultural understanding through written assignments. Some of the Indigenous languages which are taught as second languages are those which remain strong and continue to be learned by children in their home communities (e.g., Yolngu Matha at Charles Darwin University); others are languages currently undergoing processes of revitalization which are also being learned as second languages (e.g., Gamilaraay at the University of Sydney; Kuarna at the University of South Australia).

In Australia, as in Canada, school activities are managed at the state level. However, a national curriculum is currently in the process of being introduced into state schools across the country, managed by the Australian Curriculum, Assessment and Reporting Authority (ACARA). Specific content and achievement standards are specified at each year level. The ACARA National Curriculum includes a focus on Indigenous languages as a choice for schools in language education. While the curriculum documents do not detail assessment items, they do detail the aims of the courses, which include communication; understanding language, culture and learning; self-awareness within the language; and understanding language building and linguistic processes including language revitalization (Australian Curriculum, Assessment and Reporting Authority 2013, p. 5). The Curriculum also splits language learning courses into three strands: first language learner pathway for Indigenous children who are learning the language as a first language (e.g., often in remote communities); language revival learner pathway, where the language is undergoing processes of revitalization; and second language learner pathway, where there are available resources and speakers for the language to be learned as a second language in the school context. The structure of the course and subsequently the assessment are catered to suit the strand to which the course belongs.

Each state has its own syllabus documents and currently each adopts its own approach to the teaching of Aboriginal languages. In New South Wales, which has a

relatively large percentage of Aboriginal students, schools may elect to teach an Aboriginal language from K-10, and there are detailed guidelines about how to go about this. However, Aboriginal languages are not offered in the Higher School Certificate, which is taken over the last 2 years of school and contributes to evaluations for university entrance. The stated aim of the Aboriginal languages syllabus is to support the local Aboriginal communities in the revitalization of their languages, with the requirement that schools undertake widespread consultation with Aboriginal communities as well as teachers of Aboriginal languages and program managers throughout the development of the program. The K-10 syllabus assessment is designed to enhance teaching and improve learning, engaging students in "Assessment for learning" in which teachers decide how and when to assess students to ensure that assessment:

- is an essential and integrated part of teaching and learning
- reflects a belief that all students can improve
- involves setting learning goals with students
- helps students know and recognise the standards they are aiming for
- involves students in self-assessment and peer assessment
- provides feedback that helps students understand the next steps in learning and plan how to achieve them
- involves teachers, students and parents in reflecting on assessment data. (NSW Board of Studies 2003, p. 65)

The Victorian Department of Education and Training is currently working to have every student learn an Indigenous language from Prep (age 5) to year 10 (age 16). Funding for these programs is included in general school funding (State Government of Victoria Department of Education and Training 2013). The problematic aspect of this initiative is that working with a language being revitalized will cost more than a widely studied language due to lack of speakers, lack of information about the language, and lack of resources. On a national level, funding for Indigenous education is more geared towards closing the gap in terms of English literacy and numeracy rather than support of Indigenous languages (Standing Committee on Aboriginal and Torres Strait Islander Affairs 2012). However, there are schemes such as the Indigenous Language Support program (ILS), which provides funding for community groups to maintain and revive indigenous languages. This includes the creation of language learning resources (Australian Government Office for the Arts n.d.).

In South Australia, a curriculum has been developed for the language Pitjantjatjara, a central desert language spoken by around 2,000 people. Pitjantjatjara is taught in the final 2 years of schooling around three main foci: the target language, regional languages, and Australian languages. Within each of these there are subtopics that fall under *Understanding Language* or *Understanding Culture*. The final assessment for this subject is based on five capabilities: communication; citizenship; personal development; work and learning (with the aim on developing speaking, listening, reading and writing skills as well as understanding language systems); and increasing intercultural understandings (School of Languages 2015).

In the Australian school system, the focus is on learning outcomes, so while positive feedback is encouraged, quite clear direction is provided about the purpose of assessment:

> [F]eedback that students receive from completing assessment activities will help teachers and students decide whether they are ready for the next phase of learning or whether they need further learning experiences to consolidate their knowledge, understanding and skills. Teachers should consider the effect that assessment and feedback have on student motivation and self-esteem, and the importance of the active involvement of students in their own learning. (Board of Studies New South Wales 2003, p. 17)

Other states acknowledge the need for community influence and input during program design – e.g., the Northern Territory (Northern Territory Department of Education n.d.) and Victoria (Victorian Curriculum and Assessment Authority 2009) – although Victoria is less specific about community input in the creation and evaluation of assessment.

The Australian National Curriculum encourages self-assessment and peer-assessment, but there is only minimal opportunity for students to guide their own learning in an autonomous manner. Most of the state syllabi offer students the opportunity to select a certain assessment type in order to be assessed on a certain skill. For example, in the South Australian syllabus, students may choose to complete a data collection assignment and present it in a written, oral, or multimodal format (South Australian Certificate of Education 2015).

## Assessing Indigenous Children in Bilingual Schools

In the Northern Territory, where close to 30% of the population is Indigenous and often living in remote areas, there has been recognition of the benefits of bilingual schooling for children attending school with a language other than English. At the same time, bilingual schooling has been seriously challenged by frequent and rapid changes of policy (see, e.g., Simpson et al. 2009; Wigglesworth and Lasagabaster 2011). These changes have included a "first four hours of English" policy mandated in 2008. This policy resulted in many bilingual schools turning to English-only teaching, since bilingual teaching is not viable where English must be taught for the first 4 h. While a few schools still teach largely bilingual programs in which children learn literacy in their L1 for the first 4 years while acquiring English orally, the lack of available materials and of fully qualified teachers present a challenge. In addition, assessment of developing literacy in the Indigenous languages is difficult because the Northern Territory Department of Education has not developed benchmarks for the assessment of the language or curriculum (Simpson et al. 2009), and therefore tends to be informal classroom assessment.

## Indigenous Language Revitalization Activities and Assessment in Australia

There is now a considerable number of Indigenous revitalization programs in Australia, involving community members, linguists, and language centers, many of which have been recently reported on by Hobson et al. (2010). References to

assessment practices in these activities are limited. The Resource Network for Linguistic Diversity (RNLD) assists in revitalization work by providing a program (called Documenting and Revitalizing Indigenous Languages) involving a range of activities including workshops, but the program is not formally assessed. RNLD also runs a Certificate II Language Learning Program with a competency based assessment requiring participants to develop their digital literacy skills through an immersion-based Master-Apprentice framework.

The Master-Apprentice approach to language revitalization is relatively new to Australia, with only one well-established program, the Mirima Dawang Woorlabgerring Language and Culture Centre in Kununurra, Western Australia, although it is likely that the workshops being held will result in further programs (Marmion et al. 2014). Olawsky (2013), in detailing the program, identifies a number of approaches to assessment. Olawsky (2013) points out that in finding speakers to act as the Master, they needed to be fluent. At the same time, however, it was felt that it would not be appropriate for senior speakers of Miriwoong to be formally assessed. As a result, fluency was "defined as a relative criterion based on a combination of self-evaluation, judgment by other senior speakers, and linguists' experience in working with the speakers" (p. 47). As the program expanded, assessment strategies were developed for apprentices. Because the program is orally based, meaning literacy levels are not involved, the assessments were conducted in the form of a recorded interview with a linguist subsequently grading the recording.

## University-Community Partnerships in Language Assessment

Both Canada and Australia have seen a recent increase in community-based collaborations with university researchers. Academic work on language assessment in Canada has focused on the development of culturally relevant assessment tools. Morris and MacKenzie (2013) developed assessment tasks to measure lexical and morphosyntactic knowledge in child speakers in three Innu communities (in Northern Quebec and Labrador). They outline their challenges in using linguistic tools developed for non-aboriginal languages (such as word frequency counts).

Miller (2004a, b) worked with public as well as band-operated schools (schools which are managed by the local reserve) to develop "a more refined, culturally appropriate and easily administered assessment tool to determine First Nations language proficiency" (pp. 8–9). In creating a First Nations Language Benchmarks document (inspired in large part by the Canadian Language Benchmarks), stakeholder feedback led to the elimination of less culturally appropriate descriptors (such as an emphasis on the demonstrating the ability to persuade others).

Jacobs and his colleagues (2015) are currently developing a First Nations language assessment tool which is designed to work within a Mentor-Apprentice (MAP) model. Focusing primarily on listening and speaking, this tool was designed through a combination of insights from Indigenous second language learning (ISLL), decolonization theory, and sociocultural second language acquisition (SLA) theory. In MAP, learners drive the content of learning rather than using a

predetermined curriculum, so the tool is designed to track progress with blank "I CAN. . ." statement fields that can be individualized. The tool also includes opportunities for positive feedback, and has a practical graphical interface "appealing to Indigenous worldviews of information presentation" (Jacobs et al. 2015, n.d.).

On the other side of the country, Germain and Baker (2016) use a narrative inquiry methodology to examine the formative assessment practices of Germain's nature-based Mi'gmaq kindergarten immersion classroom. This exploration reveals how Germain's observations of student language use in the classroom and on the land inform her decision-making and the evaluation of her explicit and implicit learning objectives. Critical moments of assessment were identified which reveal the importance placed on the use of language for reinforcing specific community values, such as skills on the land, teamwork, and demonstration of respect for Elders.

The Canadian researchers all explicitly state their objective to affect policy making with their work. Jacobs et al. state that in addition to creating a resource for teachers, their assessment would "exercise influence on policy making related to adult Indigenous language learning," as well as "present these learners as a viable group contributing to the revival of Indigenous languages in Canada – and elsewhere" (2015, n.d.). Morris and MacKenzie (2013) note that in addition to its uses in developing pedagogy, their Innu language test data can be used "to inform language of education decisions, [and] to strengthen applications for language maintenance funding" (p. 171).

A major focus of university-based linguistic work in Australia is on the documentation of highly endangered Indigenous languages, as attested by numerous publications and theses. In terms of assessment, McConvell (1994) provided an early model for the assessment of Indigenous languages through the development of an instrument developed for Kija, an endangered language spoken in the north of Western Australia. The instrument was designed around a series of tasks which involved simple instruction, as well as questions, which could be answered either nonverbally or without a fluent response, meaning that language knowledge would not necessarily be underestimated (as would be the case where a fluent response was required).

More recent approaches to Indigenous language assessment are reported in Hudson et al. (2010). In particular, Yunkaporta (2010) discusses the importance of linking cultural and language knowledge, because teaching about "the languages of the land" [and the] "link to land and country should always be present [to ensure] cultural integrity" (Yunkaporta 2010, p. 76). He reports the following:

> In one school in western NSW some students created a sand painting using Aboriginal symbols taught by a local Elder. Another group made a story map from a local Dreaming story, using both pictures and words to show where the main incidents in the story occurred on country. Later a group of Stage 4 Aboriginal language students studied these images, linking them to the appropriate words and story in language. They then made message sticks about a common theme using those images and others to represent language words and cultural concepts based on the theme of the unit. For oral assessment they were expected to 'read' the symbols on the message sticks to the class using only the language words they had learnt. (Yunkaporta 2010, p. 76)

In the same volume, Cippolone (2010) details the development of three nationally accredited qualifications in Indigenous languages, which necessarily include assessment. Part of the assessment involved a 2 day workshop designed to recognize the prior learning of participants who had previously studied the language. Prior learning was assessed through challenge test items which were validated by experienced teachers of second languages who had specialist qualifications in applied linguistics.

Many Indigenous children who are no longer learning their traditional languages as their first language are now speaking a creole, known in Australia as Kriol, which is English lexified. Two recent studies have investigated the extent to which Indigenous children still recognize and understand words from their traditional language in an attempt to evaluate the children's potential passive knowledge of their languages. Jones and Campbell (2008) argue that assessing children's receptive rather than productive skills may provide a more accurate picture of children's knowledge of the language since their production may be hampered by the limited input to which they have access. Meakins and Wigglesworth (2013) evaluated the relationship between the input Gurindji children (who now learn Gurindji Kriol as their first language) received and their comprehension of a series of vocabulary items. The test included items which had different levels of frequency in community use, which was a critical variable in the children's comprehension. The assessment was based on a test developed by Loakes et al. (2012) which documents the challenges of developing and evaluating a suitable vocabulary assessment. These pilot results also pointed to the importance of frequency in the input.

## Conclusion

In both the Australian and Canadian contexts, we found that university researchers can be useful allies in developing culturally relevant assessment tools and in promoting the importance of recent language revitalization efforts. In addition, Indigenous language assessment practices – like pedagogical practices – are becoming more culturally relevant and useful. We found less evidence of this in university-based language courses, but school curricula are increasingly acknowledging the following characteristics of an Indigenous approach to assessment:

- The primacy of oral communication
- The importance of collaboration rather than competition, for example, through peer assessment and collaboratively-produced assessments
- The importance of involving the community in the assessment process

This last characteristic is highlighted here. In Canada, many documents include the expectation that Elders and other community members will come into the classroom and participate in assessment. For example, Alberta Education documents (2009) recommend the creation of a community-based marking method in which students self-evaluate in collaboration with peers, the teacher, and an Elder. The

documents state that "[i]t is advisable to involve Kihtêyâyak [Elders] or respected community members whenever possible in the language evaluation process" (p. 36). Similarly, in Australia, the Queensland curriculum documents signify the importance of language learning as part of the language community and acknowledge that languages belong to the communities, and that these communities, "can define their Aboriginal and Torres Strait Islander protocols and processes for their languages and knowledge," which, in combination with community ideals, "need to be the foundation upon which Aboriginal and Torres Strait Islander curriculum is created in this syllabus." The curriculum states that the syllabus "creates a space for Aboriginal and Torres Strait Islander communities to self-define the terms of entry, engagement and exit for schools seeking to meaningfully and mutually inquire into their knowledge" (Queensland Studies Authority 2010, p. 4).

In both the Australian and Canadian contexts, while Indigenous language revitalization is attracting increased attention, many challenges remain, not the least being effectively communicating to non-Indigenous educators, policy makers, and other audiences about Indigenous approaches to language assessment. The recent work of Peter Jacobs and his colleagues (2015) calls attention to this conflict between "[w]estern notions of what constitutes progress, what is success, what is valuable about what aspect of learning a language, and. . .more holistic understandings of learning and knowledge found in Indigenous worldviews" (Jacobs et al. 2015, n.d.). These conflicts must be made salient to all stakeholders and addressed with innovative pedagogies and accompanying assessment practices.

## Cross-References


- ▶ Assessing English Language Proficiency in the United States
- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ Task and Performance-Based Assessment


## Related Articles in the Encyclopedia of Language and Education


Joseph Lo Bianco, Yvette Slaughter: Bilingual Education in Australia. In Volume: Bilingual and Multilingual Education.

Joseph Lo Bianco, Yvette Slaughter: Language Policy and Education in Australia. In Volume: Language Policy and Political Issues in Education.

Fred Genesee, Joseph Dicks: Bilingual Education in Canada. In Volume: Bilingual and Multilingual Education.

Leanne Hinton: Learning and Teaching Endangered Indigenous Languages. In Volume: Second and Foreign Language Education.

Teresa McCarty, Serafin Coronel-Molina: Language Education Planning and Policies by and for Indigenous Peoples. In Volume: Language Policy and Political Issues in Education.

Onowa McIvor, Teresa L. McCarty: Indigenous Bilingual and Revitalization-Immersion Education in Canada and the United States. In Volume: Bilingual and Multilingual Education.

Donna Patrick: Language Policy and Education in Canada. In Volume: Language Policy and Political Issues in Education.

Diane Pesco, Martha Crago: Language Socialization in Canadian Indigenous Com munities. In Volume: Language Socialization.

Inge Sichra: Language Diversity and Indigenous Literacy in the Andes. In Volume: Literacies and Language Education.

Sabine Siekmann: Indigenous Language Education and Uses of Technology. In Volume: Language, Education and Technology.

## References

ABS (2011). Australian Social Trends March 2011: Life expectancy trends – Australia. Canberra: Australian Bureau of Statistics.

Alberta Education. (2009). *Cree language and culture 12-year program guide to implementation: Kindergarten to grade 3. Curriculum sector: Arts, communications and citizenship*. Government of Alberta

Australian Curriculum, Assessment and Reporting Authority. (2013). *Australian curriculum: Languages foundation to year 10 draft framework for Aboriginal languages and Torres Strait Islanders languages*. Retrieved from http://consultation.australiancurriculum.edu.au/Static/docs/Languages/Aboriginal%20Languages%20and%20Torres%20Strait%20Islander%20Languages%20Draft%20Framework%20%20-%20May%202013.pdf

Australian Government, Office for the Arts. (n.d.). Indigenous languages support. Retrieved from http://arts.gov.au/sites/default/files/indigenous/ils/ils-factsheet.pdf

Board of Studies New South Wales. (2003). Aboriginal languages K-10: Syllabus. Retrieved from http://www.boardofstudies.nsw.edu.au/syllabus_sc/pdf_doc/ab_language_k10_syl.pdf

Cippolone, J. (2010). Aboriginal languages programs in TAFE NSW: Delivery initiatives and strategies. In J. Hobson, K. Lowe, S. Poetsch, & M. Walsh (Eds.), *Re-awakening languages: Theory and practice in the revitalisation of Australia's Indigenous languages*. Sydney: Sydney University Press.

Fagan, K. (2001). *Laughing to survive: Humour in contemporary Canadian Native literature*. Unpublished Ph.D. dissertation, University of Toronto. Retrieved from http://www.collectionscanada.gc.ca/obj/s4/f2/dsk3/ftp05/NQ63653.pdf

Germain, J., & Baker, B. (2016, June). A narrative inquiry into the formative assessment practices of an Indigenous language teacher. In *Symposium: Assessment of Indigenous languages: Examining the cases of Australia and Canada. Language testing research colloquium (International Language Testing Association annual conference)*, Palermo.

Government of Ontario. (2001). *The Ontario curriculum, grades 1–8: Native languages*. Government of Ontario

Government of the Northwest Territories, Early Childhood and School Services. (2015). *Curriculum documents grades K-9. Education: A Dene perspective*. Retrieved from http://www.ece.gov.nt.ca/early-childhood-and-school-services/school-services/curriculum-k-12/aboriginal-languages

Government of the Northwest Territories, Early Childhood and School Services. (2015). *Curriculum document K 12. Inuuqatigiit: The curriculum from the Inuit perspective*. Retrieved from http://www.ece.gov.nt.ca/early-childhood-and-school-services/school-services/curriculum-k-12/aboriginal-languages

Hudson, J., Lowe, K., Poetsch, S., & Walsh, M. (Eds.). (2010). *Re-awakening languages: Theory and practice in the revitalisation of Australia's Indigenous languages*. Sydney: Sydney University Press.

Jacobs, P., McIvor, O., Jenni, B., & Anisman, A. (2015, Mar). *NĒTOLNĒW "One mind, one people": Developing a context-relevant assessment tool for adult indigenous language learners*. Paper presented at the American Association of Applied Linguistics Annual conference, Toronto.

Jones, C., & Campbell, N. (2008). Issues in the assessment of children's oral skills. In J. H. Simpson, & G. Wigglesworth (Eds.), *Children's language and multilingualism: Indigenous language use at home and school* (pp. 175–193). London: Continuum Int Pub Group.

Kwayaciiwin Education Resource Centre. (2014a). *Kwayaciiwin curriculum general guide document*. Sioux Lookout.

Kwayaciiwin Education Resource Centre. (2014b). *Immersion guide document*. Sioux Lookout.

Langlois, S., & Turner, A. (2013). Aboriginal languages in Canada in 2011. In M. J. Norris, D. Patrick, & N. Ostler (Eds.), *Proceedings for the seventeenth conference of the foundation for endangered languages* (pp. 162–169). Ottawa: Carleton University.

Loakes, D., Moses, K., Simpson, J., & Wigglesworth, G. (2012). Developing tests for the assessment of traditional language skill: A case study in an Indigenous Australian community. *Language Assessment Quarterly, 9*(4), 311–330.

Manitoba Education, Citizenship, and Youth. (2007). *Kindergarten to grade 12 Aboriginal anguages and cultures: Manitoba curriculum*. Winnipeg: Government of Manitoba.

Marmion, D., Obata, K., & Troy, J. (2014). Community, identity, wellbeing: The report of the Second National Indigenous Languages Survey. Canberra: Australian Institute of Aboriginal and Torres Strait Islander Studies.

McConvell, P. (1994). Oral proficiency assessment for Aboriginal languages. In D. Hartman & J. Henderson (Eds.), Aboriginal Languages in Education (pp. 301–15). Alice Springs: IAD Press.

McKay, G. (2011). Policy and Indigenous languages in Australia. *Australian Review of Applied Linguistics, 34*(3), 297–319.

Meakins, F., & Wigglesworth, G. (2013). How much input is enough? Correlating passive knowledge and child language input. *Journal of Multilingual and Multicultural Development, 34*(2), 171–188.

Miller, J. W. (2004a). *A language teacher's guide to assessing first Nations language proficiency*. Retrieved from http://www.interiorsalish.com/languageassessment.html

Miller, J. W. (2004b). *Assessing first Nations language proficiency*. Unpublished Ph.D. dissertation, University of Victoria, Canada.

Morris, L., & MacKenzie, M. (2013). Using all the pieces to solve the puzzle: The importance of Aboriginal language assessment in child populations. In M. J. Norris, D. Patrick, & N. Ostler (Eds.), *Proceedings for the seventeenth conference of the foundation for endangered languages* (pp. 170–177). Ottawa: Carleton University.

Northern Territory Department of Education. (n.d.). *Indigenous languages and culture*. Retrieved from http://www.education.nt.gov.au/__data/assets/pdf_file/0014/2372/indigenous_lang_cult.pdf

Olawsky, K. J. (2013). The Master-Apprentice Language Learning Program Down Under: Experience and adaptation in an Australian context. Language Documentation and Conservation, 7:41–63.

Patrick, D. (2010). Canada. In J. Fishman & O. Garcia (Eds.), *Handbook of language and ethnic identity* (2nd ed., pp. 286–301). Oxford: Oxford University Press.

Queensland Studies Authority. (2010). *Aboriginal and Torres Strait Islander languages: P-10 Queensland syllabus 2010*. Retrieved from https://www.qcaa.qld.edu.au/downloads/p_10/atsi_languages_P-10_syll.pdf

Research Network for Linguistic Diversity (RNLD). (n.d.). Retrieved from http://www.rnld.org/CertII

School of Languages. (2015). *SACE subject overview 2015*. Retrieved from http://www.schooloflanguages.sa.edu.au/documents/PitjantjatjaraS1-S2courseOverview2015.pdf

Simpson, J., Caffery, J., & McConvell, P. (2009). *Gaps in Australia's Indigenous language policy: Dismantling bilingual education in the Northern Territory* (Australian Institute of Aboriginal & Torres Strait Islander Studies Research Discussion Paper, 23). Canberra: Aboriginal Studies Press [Online]. Retrieved from www.aiatsis.gov.au/research/docs/dp/DP24.pdf

South Australian Certificate of Education. (2015). *Australian languages*. Retrieved from https://www.sace.sa.edu.au/web/australian-languages

Spielmann, R. (1998). *"You're so fat!" Exploring Ojibwe discourse*. Toronto: University of Toronto Press.

Standing Committee on Aboriginal and Torres Strait Islander Affairs. (2012). *Our Land Our Languages*. Retrieved from http://www.aph.gov.au/Parliamentary_Business/Committees/House_of_representatives_Committees?url=/atsia/languages2/report.htm

State Government of Victoria Department of Education and Training. (2013). *Policy, government school funding and curriculum*. Retrieved from http://www.education.vic.gov.au/school/teachers/teachingresources/discipline/languages/pages/policy.aspx

The Western Canadian Protocol for Collaboration in Basic Education, Kindergarten to Grade 12. (2000). *The common curriculum framework for aboriginal language and culture programs: Kindergarten to grade 12*. Retrieved from https://www.wncp.ca/english/subjectarea/fnmi/commoncurriculumframework.aspx

Truth and Reconciliation Commission of Canada. (2015). *Honouring the truth and reconciling for the future: Summary of the final report of the Truth and Reconciliation Commission of Canada*. Winnipeg: Truth and Reconcilliation Commission of Canada.

UNESCO. (2011). *UNESCO Atlas of the world's languages in danger*. Retrieved from http://www.unesco.org/culture/languages-atlas/index.php

Victorian Curriculum and Assessment Authority. (2009). *Aboriginal languages, cultures and reclamation in Victorian schools: Standards P-10 and protocols*. Retrieved from http://www.vcaa.vic.edu.au/Documents/vels_aboriginal_languages.pdf

Wigglesworth, G., & Lasagabaster, D. (2011). Indigenous languages, bilingual education and English in Australia. In C. Norrby & J. Hajek (Eds.), *Uniformity and diversity in language policy* (pp. 141–156). Bristol: Multilingual Matters.

Wilson, J. W. (2004). Assessing first nations language proficiency. Unpublished PhD Dissertation, University of British Colombia, Canada.

Yunkaporta, T. K. (2010). Our ways of learning Aboriginal languages. In J. Hobson, K. Lowe, S. Poetsch, & M. Walsh (Eds.), *Re-awakening languages: Theory and practice in the revitalisation of Australia's Indigenous languages*. Sydney: Sydney University Press.

# Utilizing Accommodations in Assessment

Jamal Abedi

**Abstract**

English language learners (ELLs) usually perform lower than native English speakers academically mainly due to the language factors. A majority of ELLs have the content knowledge but are not at the level of English proficiency to understand teacher's instruction and test questions. The longer ELL students stay in English-only (EO) academic environment, the smaller the performance gap becomes between ELLs and EOs. Therefore, to have equal and fair educational opportunity for everyone, ELLs must be provided with tools to help them overcome their language difficulties. These tools often refer to as accommodations. For example in mathematics assessment, providing glossaries of complex English terms unrelated to mathematics content or providing customized dictionaries where content-related terms are removed are two examples of language-based accommodations for ELL. Similarly, students with disabilities (SWD) need accommodations to help them with their disabilities. For example, students who are hard-of-hearing need hearing aids to deal with their hearing problems. However, accommodations that are provided for ELLs and SWDs should only help them deal with limited language proficiency (for ELLs) and disabilities, not to provide unfair advantage to the recipients. If they do, then the accommodated assessment outcomes will not be valid. Therefore, effectiveness and validity are two important characteristics of all forms of accommodations. An accommodation is effective if it helps remove the construct-irrelevant sources and make assessments more accessible for the recipients and is valid if it does not alter the focal construct.

J. Abedi (✉)
School of Education, University of California, Davis, CA, USA
e-mail: jabedi@ucdavis.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_21

**Keywords**

Accommodation • Assessment • English language learners • Students with disabilities • Effectiveness • Validity

## Contents



## Introduction

A fair assessment and accountability system in many countries requires that all students be included in large-scale national and local assessments. However, research clearly demonstrates a substantial performance gap between those for whom the assessment language is a second language and those students who are native speakers of the assessment language, particularly on academic subjects that are high in language demand (Abedi 2006a). The literature suggests that this performance gap is explained by many different factors including parent education level and support, SES, the challenge of second language acquisition (Hakuta et al. 2000; Moore and Redd 2002), and a host of inequitable schooling conditions (Gándara et al. 2003). Yet, it is also often the case that the measurement tools are ill-equipped to assess the skills and abilities of second language learners. To offset these challenges, nonnative speakers of the assessment language are provided with "test accommodations."

Accommodations are used to make assessments more accessible for English language learners and students with disabilities and to produce results that are reliable and valid for these students without altering the focal construct (Abedi and Ewers 2013).

Test accommodations refer to changes in the test process, in the test itself, or in the test response format. The goal of accommodations is to provide a fair opportunity for nonnative speakers of the assessment language and students with disabilities to demonstrate what they know and can do, to level the playing field, so to speak, without giving them an advantage over students who do not receive the accommodation.

The issues concerning accommodations are important in all countries where there are students who do not have high proficiency in the language of instruction and assessment in schools; usually these are immigrants and indigenous groups. Since researchers in the USA have conducted more research on accommodations than many other countries, in this chapter we present an overview of major research findings that are reported in the American research journals for English language learners (ELLs).

Literature has clearly demonstrated that there are many accommodations currently used for ELLs and students with disabilities. However, care must be exercised in selecting appropriate accommodations. The literature suggests that many of the accommodations created and used for students with disabilities are used for ELLs without any evidence of effectiveness of these accommodations for this group of students. ELL students need language-based accommodations to facilitate their understanding of teachers' instruction and language of test items (Abedi 2012; Abedi and Ewers 2013).

To be useful and to provide reliable and valid assessment results, accommodations must meet the following major conditions (for a more detailed discussion of these conditions see Abedi 2012):

Based on extensive literature review on accommodations for ELLs and students with disabilities (SWDs), Abedi and Ewers (2013) provide five important conditions under which accommodations can be validly used for ELLs and SWDs. These conditions include: (1) Effectiveness, (2) Validity, (3) Differential Impact, (4) Relevance, and (5) Feasibility. Below is a short description of each of these five conditions.

## Effectiveness

Accommodations must be effective in making assessment more accessible to the recipients by controlling for the construct-irrelevant factors. For example, ELL students need assistance in language of instruction and language of assessments particularly when the language is unnecessarily complex. Accommodations such as glossary of uncommon or difficult vocabulary, native language assessment, and customized dictionaries would be quite helpful and effective. For example, Li and Suen (2012) using a meta-analysis approach found some of these accommodations improved performance of ELL students while not impacting the performance of non-ELLs. However, the authors indicated that the level of impact on ELLs was not substantial (0.156 standard deviation). Wolf and her colleagues (2012) examined the effectiveness of read-aloud and glossary accommodations in making assessments more accessible for ELL students. The authors found no significant impact of including a glossary, but they found some impact of read-aloud strategies and significant interaction between students' prior knowledge and accommodations (see, also Willner et al. 2009).

## Validity

An accommodation can be valid if it does not alter the focal construct or provide unfair advantage to the recipients so that the outcomes of accommodated and nonaccommodated assessments should be comparable and combinable. Invalid accommodations affect the outcome of assessments for individual students as well as for the group in which students belong. If accommodations affect the construct, then the accommodated and nonaccommodated assessments cannot be aggregated. Studies have found that some forms of accommodations may alter the construct being measured (see, for example, Abedi et al. 2004). For example, providing a published dictionary may affect the measurement of the construct, since it may provide content-related information which students can use to answer the questions. Abedi et al. (2004) found that providing a glossary plus extra time increased performance of non-ELL students for whom the accommodation was not intended, thereby increasing the performance gap between ELL and non-ELL students. Thus, the validity of many commonly used accommodations is questionable. Unfortunately, research on the validity of accommodations is very limited and the validity of only a handful of accommodation strategies used for ELL students have been experimentally examined (Abedi et al. 2004; Francis et al. 2006; Sireci et al. 2003).

The best and the most straight approach in examining the validity of accommodation is to randomly assign non-ELL students (who do not need accommodations) to a treatment group where they are tested under an accommodation or to control group where they are tested under standard condition without any accommodations. If non-ELLs who are tested under the accommodation perform significantly different under the accommodated condition, then the accommodation does more than what is supposed to do, i.e., it changes the focal construct. Kieffer et al. (2009) examined the effectiveness and validity of accommodations using a meta-analytical approach. They found that none of the seven accommodations used in this study alter the focal construct, therefore, they can be used without being concerned about the validity of these accommodations.

## Differential Impact

To be effective and useful, an accommodation should fit with student's background characteristics and their academic standing, i.e., one size may not fit all. The ELL population is quite diverse and consists of students with very different academic, personal, and family backgrounds. For example, they are different in terms of their proficiency in their native (L1) and English (L2) languages. Some of them are quite proficient in both L1 and L2, some are more proficient in one, and some are not proficient in either. They are also different in their levels of content knowledge. For example, Wolf et al. (2012) found that some of the accommodations they used are more effective for students with content knowledge.

## Relevance

Accommodations used for ELLs should be consistent with their academic backgrounds and needs. Many accommodations that are used for ELLs were initially created for and used by students with disabilities. Later in this chapter, we will provide examples of these accommodations. The most relevant accommodations for ELL students are language-based accommodations because that is what ELL students need. For example, ELL students have difficulty with the test items that have complex linguistic structure; therefore, accommodations such as providing glossaries of noncontent terms and customized dictionaries would be more relevant.

## Feasibility

An accommodation must be logistically feasible to implement during assessments. Accommodations that are effective and provide valid results may be selected because of difficulty in administration. For example, one of the accommodations used for ELLs are providing them with commercial dictionaries. There are two ways this accommodation is implemented, either students bring their own dictionary or a standard one is provided to them by test administrators. The first option adds a construct-irrelevant factor, which are the differences between individual students' dictionaries. The second option has its own problem of delivering and collecting dictionaries. Furthermore, another example we can discuss is the application of computer testing, which could be a burden if a school lacks funding for adequate computer resources. One-on-one testing may also be logistically challenging in large-scale assessments.

Accommodations that meet all the five requirements discussed above, particularly effectiveness and validity, can provide assessments that are more accessible for ELLs and students with disabilities (SWDs) without altering the focal construct. Such accommodations may also be considered for all students as *accessibility features* because they control for sources of construct-irrelevant factors.

As briefly mentioned above, the effectiveness, validity, and differential impact of accommodations can be examined through a randomized controlled trial (RCT) experiment in which most of the accommodations can be randomly assigned to students. As such, sources of threats to internal and external validity of the experiment can be controlled. In this experiment, ELL and non-ELL students are randomly assigned to the accommodated and nonaccommodated conditions that allows for the examination of effectiveness and validity. For example, if accommodated ELL students performed better than nonaccommodated ELLs in a content area such as mathematics, then the accommodation is considered as "effective." On the other hand, if non-ELLs under accommodation perform higher than non-ELLs who are tested under the standard testing condition with no accommodation, then the accommodation is believed to have altered the focal construct.

## Early Developments

Historically, the concept of accommodations was first introduced in the field of special education. Many students with disabilities need specific forms of assistance in the classroom setting to deal with their disabilities, i.e., to level the playing field. For example, deaf and hard-of-hearing students need hearing aids to offset the effect of their inability to hear at the same level as regular students. Similarly, blind or visually impaired students need to use the brail version of a test or vision aids to be able to read the test items. These accommodations are used to increase equity in the classroom as well as in assessments. The concept of accommodations was then extended to ELLs. Unfortunately, however, not only the concept of accommodations but also the strategies that were created and used for students with disabilities were used for ELL students, many of which may not be relevant for these students.

By definition, accommodations are used for students with disabilities (SD) to assist them with their disabilities. For ELL and nonnative speakers of the assessment language the goal of accommodations is to help with second language needs. Another goal is to reduce the performance gap between SD/ELL and non-SD/non-ELL students, without jeopardizing the validity of assessments. In the USA, there are many forms of accommodations which are used for both ELL students and students with disabilities in different states (Abedi et al. 2000; Rivera et al. 2000; Thurlow and Bolt 2001). Yet, as will be shown below, there is little evidence to support the effectiveness and validity of assessments using these accommodations.

## Major Contributions

As noted above, the main focus of this chapter is on accommodations for ELL students in the USA. However, a short discussion on accommodations for students with disabilities must be included as well, due to some historical connections between the accommodation policies and practices for these two subgroups of students. In fact, some accommodations that are currently used for ELL students were initially developed and used for students with disabilities (see, for example, Rivera et al. 2000).

Review of literature on accommodations suggests that: (1) existing research on some forms of accommodations is not conclusive, and (2) for many forms of accommodations used by different states there is very limited empirical data to support their validity. It should be noted that the term "validity of accommodations" is used here within the general framework of assessment; therefore, validity of accommodations refers to the "validity of accommodated assessments." In other words, an accommodation strategy may not be valid or invalid unless it is considered within the assessment framework. In presenting the research summary it will be shown that: (1) some accommodations that are used for ELL students are designed for students with disabilities and are not relevant to ELL students, and (2) in some cases, findings from different studies about accommodations are not consistent. Below is a summary of research for some commonly used accommodations.

Braille is used for students with blindness or significant visual impairments. Braille versions of a test may be more difficult for some items than other items such as items with diagrams and/or special symbols (Bennet et al. 1987b, 1989; Coleman 1990). This is clearly an accommodation for SD (blind) students only.

Recently two consortia of states (Smarter Balanced and the Partnership for Assessment of Readiness for College and Careers [PARCC]) developed computer-based assessments, and states will shift from paper-and-pencil mode to computer-administered assessments. Computerized Assessment is especially helpful for students with physical impairments that have difficulty in responding to items in a paper-and-pencil format. Some studies suggest that this accommodation increases the performance of students (Russell 1999; Russell and Haney 1997; Russell and Plati 2001). Other studies have not found computerized assessments to be effective (Mac Arthur and Graham 1987), or not as effective as traditional assessments (Hollenbeck et al. 1999; Varnhagen and Gerber 1984; Watkins and Kush 1988). In a study with grade 4 and 8 students in mathematics, Abedi et al. (see, Abedi et al. 2004) found that computerized assessments can be highly effective in making tests more accessible to ELL students. The study did not find any validity issues with the computerized assessment suggesting that the computerized assessment did not impact the assessment of focal construct.

Dictate Response to a Scribe (someone writes down what a student dictates with an assistive communication device). This accommodation has been shown to have an impact on the performance of students with learning disabilities (Fuchs et al. 2000; Mac Arthur and Graham 1987). Tippets and Michaels (1997) found this accommodation, in combination with other accommodations, such as read aloud and extended test time helps students with disabilities. However, there are concerns over the validity of this accommodation. Koretz (1997) found this accommodation helped students with learning disabilities; however, Thurlow and Bolt (2001) recommended that if students are unable to handwrite but can efficiently use a computer, the use of a computer should be considered.

Extended Time. This is one of the most commonly used accommodations. Under this accommodation, students receive extra time (usually 50% more time) to respond to the test items. It is used for both English language learners and students with different types of disabilities. Thurlow et al. (2000) suggested that disagreement between states may be a concern regarding the validity of extended time accommodation. Chiu and Pearson (1999) found extended time to be an effective accommodation for students with disabilities, particularly for learning disabilities. Some studies found extended time to help students with disabilities in Mathematics (Chiu and Pearson 1999; Gallina 1989). However, other studies did not show an effect of extended time on students with disabilities (Fuchs et al. 2000; Marquart 2000; Munger and Loyd 1991). Studies on the effect of extended time in language arts did not find this accommodation to be effective (Fuchs et al. 2000; Munger and Loyd 1991). Some research studies showed that extended time affects the performance of both SD and non-SD students, and therefore makes the validity of this accommodation suspect. For ELL students, research on extended time has produced mixed results. Abedi et al. (2004) found no effect of extended time for ELL students.

On the other hand, Hafner (2000) found extended time to be an effective accommodation for ELL students.

It must be noted at this point that many school districts in the USA allow unlimited time in taking both Title I and Title III assessments (Rivera and Collum 2006) under the No Child Left Behind (NCLB 2001) accountability requirements. That is, the state tests are often considered as power tests and not as speed tests. Therefore, extended time is not viewed as an accommodation and consequently there is no concern over the validity of assessments using extended time since everyone receives extra time in testing.

Interpreter for Instructions. In this accommodation an interpreter translates test instructions in sign language. This accommodation is recommended for students with hearing impairments. Adaptations in the presentation of directions may help deaf children score the same as other students (Sullivan 1982).

Large Print is used for students with visual impairments. Research has indicated that this accommodation has helped reduce the performance gap between students with visual impairments and students without disabilities (see, for example, Bennet et al. 1987a). The results of a study by Bennet et al. (1987b) revealed that using this accommodation for visually impaired students does not affect the construct under measurement. Other studies suggest that extra time may be needed with this accommodation (Wright and Wendler 1994). Large print has also been used for students with learning disabilities. Several studies have shown no impact of this accommodation for students with learning disabilities. One study, however, showed that large print helps students with learning disabilities (Perez 1980). This accommodation has also been used for ELL students (Rivera 2003; Sireci et al. 2003) although it is not clear how relevant this accommodation is to ELL students.

Allowing students to mark answers in test booklet, rather than on an answer sheet is another commonly used accommodation. This accommodation can be used for students who have a mobility coordination problem. Some studies on the effectiveness of this accommodation did not find significant difference between those students tested under this accommodation and those using separate answer sheets (Rogers 1983; Tindal et al. 1998). However, other studies found lower performance for students using this accommodation (Mick 1989). In fact, many school districts in the USA have used this accommodation for ELL students (Rivera 2003), yet there is no evidence on the relevance or effectiveness of this accommodation for ELL students.

Read Aloud Test Items are used by students with learning disabilities and students with physical or visual impairments. While some studies found this accommodation to be valid in mathematics assessments (Tindal et al. 1998), others have concerns over the use of this accommodation on reading and listening comprehension tests (see, for example, Burns 1998; Phillips 1994) since this accommodation may impact the validity of assessment by altering the construct (see also Bielinski et al. 2001; Meloy et al. 2000). Read aloud as an accommodation has also been used for ELL students in the USA (Rivera 2003), again, without any indication of the relevance or effectiveness of this accommodation for this group of students.

Reading or Simplifying Test Directions is appropriate for students with reading/learning disabilities. A study by Elliot et al. (2001) suggested that this accommodation affects performance of both students with disabilities and students without disabilities. There are therefore concerns over the validity of this accommodation especially since it has also been used frequently for ELL students; the use of this accommodation is of particular concern in reading assessment.

Test Breaks where students receive multiple breaks during the testing session can help students with different forms of disabilities. A study by DiCerbo et al. (2001) found that students tested under the multiple-breaks administrations obtained significantly higher scores than those tested under standard testing conditions with no additional breaks. The study also showed that middle and low-ability readers benefited more from this accommodation than high-ability readers. However, another study (Walz et al. 2000) found that students with disabilities did not benefit from a multiple-breaks test administration while students without disabilities did. These results show quite the opposite of what is expected of valid accommodations. Sometimes test breaks as a form of accommodation has been recommended for ELL students (Rivera 2003) as it may help some ELL students but may not be relevant for other ELLs since it does not address their English language needs.

Providing an English dictionary and extra time (Abedi et al. 2004; Hafner 2000; Thurlow 2001) was found to affect performance of all students (see also, Maihoff 2002; Thurlow and Liu 2001). This suggests that the results of accommodated and nonaccommodated assessment may not be aggregated.

Translation of Assessment Tools into Students' Native Language may not produce desirable results and may even provide invalid assessment results if the language of instruction and assessment is not aligned (Abedi et al. 2004).

As noted earlier, in spite of the concerns expressed by researchers over the validity, effectiveness, and feasibility of some forms of accommodations, these accommodations are used frequently by states and districts across the USA. That is, decisions on the type of accommodations for English language learners and students with disabilities do not seem to have been influenced much by the research findings.

Accommodation Issues for English Language Learners: Accommodations are meant to "level the playing field" for ELL students by accommodating their potential language limitations in an assessment. Unfortunately, there are major equity issues with many of the accommodations used for ELL students. The practice of using accommodations for ELL students that are initially developed for students with disabilities (Rivera et al. 2000) is extremely problematic as some accommodations that are used for students with disabilities are not relevant for ELL students. For example, using large print may be an effective accommodation for some students with visual impairments while ELL students need specific accommodations to address their linguistic needs. As discussed above, there are major issues concerning accommodations for both ELLs and students with disabilities. While these issues

deserve equal attention for both SD and ELL students, the focus in the next section will be on accommodation issues for ELL students.

## Work in Progress

As the number and percentage of English language learners increase in the USA, assessment equity and validity are becoming priorities for educational policymakers. Between 1990 and 1997, the number of US residents born outside the country increased by 30%, from 19.8 million to 25.8 million (Hakuta and Beatty 2000). According to the National Clearinghouse for English Language Acquisition, over 4.5 million Limited English Proficient (LEP) students were enrolled in US public schools in 2000–2001, representing nearly 10% of the nation's total public school enrollment for prekindergarten through Grade 12 (Kindler 2002).

To reduce the impact of language factors on the assessment outcome of ELL students, assessment in students' native language has been proposed as an accommodation. While this seems to be an attractive idea and many districts and states in the USA use this approach, research results do not support its fairness (Abedi et al. 2004). One major issue here is the possibility of lack of alignment between the language of instruction and language of assessment. If the language of assessment is not the same as the language of instruction, then the assessment outcome may be even less valid, again raising fairness as a serious issue. For example, when a native Spanish speaker learns content-area terminology in English, but is tested in Spanish, the outcome of the assessment may not be valid due to the student's lack of content terminology knowledge in Spanish. A student may be a fluent speaker of a language but not necessarily proficient in the academic language of his or her native language.

Some educational researchers and policymakers suggest that rather than testing students in their native languages (L1), they should be assessed by providing them with language accommodations such as a customized dictionary or a linguistically modified version of the test to help them with their English language needs. This seems to be a reasonable approach if the focus is on learning English as quickly as possible. However, others argue that students' knowledge of their first language could benefit their academic progress, and testing them in English may not properly utilize their knowledge of L1.

## Problems and Difficulties

The purpose of testing accommodations is to assist students with certain limitations that they might have and provide them with a fair assessment. It is therefore important to examine the appropriateness, effectiveness, validity, and feasibility of accommodations for the targeted student populations.

Appropriateness. How appropriate are accommodations that are provided for ELL students? Since the common characteristic that distinguishes ELLs from non-ELL students is their possible limitation in English proficiency, it is reasonable

to expect that accommodations that help ELL students with their language barrier would be the most relevant. However, in many places, the current practice of accommodations for ELL students is to simply use accommodations that are easily available or those that decision makers find relevant. These accommodations may not always be appropriate for these students. For example, Rivera (2003) presented a list of 73 accommodations that are used nationally for ELL students. Our analyses of these accommodations (Abedi 2006b) revealed that of these 73 accommodations, only 11 (15%) of them were highly relevant for ELL students in providing assistance with students' language needs. The list included accommodations such as:

- Subtests flexibly scheduled
- Tests administered at a time of day most beneficial to test-taker
- Tests administered in small groups
- Tests administered in a familiar room
- Colored stickers or highlighters for visual cues provided
- Copying assistance provided between drafts
- Test-taker types or uses a machine to respond (e.g., typewriter/word processor/computer)
- Test-taker indicates answers by pointing or other method
- Test-taker verifies understanding of directions

Since none of these accommodations address ELL students' language needs, they may not be adequate or appropriate for these students. The National Assessment of Educational Progress (NAEP) also uses some accommodations that, at face value, are not very relevant to ELL students' language needs. For example, among the accommodations NAEP used for ELL students in the 1998 civics assessment were large print, extended time, reading questions aloud, small group testing, one-on-one testing, and scribe or computer testing (see Abedi and Hejri 2004). While some of these accommodations may be helpful for students with disabilities, they may not be effective for ELL students. Studies have found that the provision of accommodations in NAEP increased the inclusion rate of these students (Mazzeo et al. 2000). However, research has shown that accommodations did not increase ELL student scores on the NAEP; that is, providing accommodations did not reduce the performance gap between ELL and non-ELL students. For example, no statistically significant differences were found between the performance of accommodated and nonaccommodated ELL students in the 1998 NAEP main assessments in reading, writing, and civics for students in fourth and eighth grades (Abedi and Hejri 2004). Among the most likely explanations for this is the lack of relevant accommodations. As indicated earlier, if the accommodations provided to ELL students have no relevance to their needs (mainly English language proficiency), then one would not expect any positive impact of accommodations on the outcome of assessments. Examples of relevant accommodations for ELLs and nonnative speakers of the assessment language include providing a glossary of noncontent terminology or modifying complex linguistic features as these accommodations directly address ELL students' language needs.

Another major issue in the provision of accommodations in NAEP was the very small number of ELL students who were accommodated. In the main NAEP assessments, the number of ELL students who were included in the study comprised between 7% and 8% of the sampled students, but only a fraction of these students, who had been accommodated by their schools in earlier assessments, received NAEP accommodations. For example, in the main assessment of the 1998 Grade 4 reading test, 934 ELL students were included, but only 41 (4%) of them were provided with accommodations. In the Grade 8 sample, 896 ELL students were included, but only 31 (3.5%) were accommodated. Similarly, in the 1998 main assessment in civics, 332 ELL students in Grade 4 were included and only 24 (7%) were accommodated. In the same assessment, 493 ELL students were included in Grade 8, but only 31 (6%) were accommodated (Abedi and Hejri 2004).

## Future Directions

Research-supported accommodations. The main goal of an accommodation is to make assessments more accessible across subgroups of students who otherwise could be affected unfairly by many nuisance variables that would make the assessment unfair and invalid. The discussion above casts doubt over the ability of many of the current accommodation practices to reach this important goal. There is no firm evidence to suggest that the accommodations used widely by school districts are effective, feasible, and valid. However, results of recent studies introduce some accommodation strategies for ELL students that, in addition to being valid, are also effective in reducing the performance gap between ELL and non-ELL students in content-area assessments.

One major assessment issue is that a student's level of proficiency in the language of assessment may severely impact the validity of the assessment results. Students may have the content knowledge (e.g., in math and science) in their native language but may not be fluent enough in the language of assessment to express their knowledge on a test. To reduce the impact of language factors on the assessment outcomes of students, the linguistic modification of test items has been proposed in the literature (see, for example, Abedi et al. 1997). A linguistic-modification approach helps test developers reduce the level of unnecessary linguistic complexity in test items by controlling for sources of linguistic complexity (for a detailed description of linguistic modification approach, see Abedi 2006a).

Earlier in this chapter research-based evidence about accommodations was presented. This evidence raises concerns about the validity of the accommodations used in schools for ELL students. The main question for the future is whether there are accommodations that would be beneficial to ELL students but do not affect the construct under measurement. Below is a short survey of accommodations that studies have shown to be effective and valid.

Recent studies at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) have examined several different forms of accommodation. Abedi et al. (2004) and Maihoff (2002) examined the linguistic modification

approach and found it to be an effective and valid accommodation in the assessment of ELL students. Rivera and Stansfield (2001) found this accommodation to have no impact on the non-ELL student group suggesting that the accommodation is valid for ELL students. With this approach, simpler versions of items with language that might be difficult for students were drafted; the task remained the same, but noncontent vocabulary and unnecessary linguistic complexity were modified (see Abedi 2006a, for further discussion of the nature of and rationale for the linguistic modifications). These studies compared student scores on NAEP test items with comparable modified items in which the mathematics tasks and mathematics terminology were retained but the language and/or linguistic structures were modified.

Following are a few examples of studies on the effectiveness and validity of the linguistic modification approach as a form of accommodation for ELL students. Abedi and Lord (2001) examined the effects of this accommodation with 1,031 eighth grade students in southern California. Test booklets with either original English versions or modified English versions of the items were randomly assigned to the students. The results showed significant improvements in the scores of students in low- and average-level mathematics classes who received the booklets with linguistic modifications. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive voice verb constructions. English language learners and low-performing students benefited the most from the linguistic modification of test items.

In another study, Abedi et al. (2004) examined the impact of linguistic modification on the mathematics performance of English learners and non-English learners. Using items from the 1996 NAEP Grade 8 Bilingual Mathematics booklet, three different test booklets (Original English, Modified English, and Original Spanish) were randomly distributed to a sample of 1,394 eighth grade students in schools with high enrollments of Spanish speakers. Results showed that language modification of items contributed to improved performance on 49% of the items. The students generally scored higher on shorter problem statements.

A third study (Abedi et al. 2004) examined the impact of four different forms of accommodation on a sample of 946 eighth grade students tested in math. The accommodations were (1) Modified English, (2) Extra Time only, (3) Glossary only, and (4) Extra Time plus Glossary. These four accommodation types, along with a standard test condition, were randomly assigned to the sampled students. Findings suggested that some accommodations increased performance of both English learners and non-English learners, compromising the validity of the assessment. Among the different options, only the Modified English accommodation narrowed the score gap between English language learners and other students.

Other studies have also employed the language modification approach. Kiplinger et al. (2000) found linguistic modification of math items helpful in improving the math performance of ELL students. Maihoff (2002) found linguistic modification of content-based test items to be a valid and effective accommodation for ELL students in math. Rivera and Stansfield (2001) compared English language learner performance on regular and modified fourth and sixth grade science items. Although the small sample size did not show significant differences in scores, the study

demonstrated that linguistic modification did not affect the scores of English-proficient students, indicating that linguistic modification is not a threat to score comparability.

While the current prevalent trends in accommodation practices are not supported by research (Solano–Flores and Trumbull 2003), there is growing evidence that states are paying more attention to research findings on the effectiveness and validity of accommodations. The increasing use of research-supported accommodations for ELL students (such as linguistic modification of items) is encouraging. This trend may result in fairer assessments for ELL students.

English language learners and students with disabilities are faced with many challenges in their academic career and need special attention. For ELL students, the challenge of learning English and at the same time competing with their native English speaking peers in learning academic concepts in English is enormous. Similarly, for students with disabilities, it is quite challenging to learn at the same rate as their nondisabled peers given their disabilities. Even more serious is the case of ELL students with disabilities. These students are faced with dual challenges – learning a new language and dealing with their disabilities. Such inequity in educational opportunity creates a substantial performance gap between these students and their peers. While accommodations are provided to offset these challenges, it has been shown that these accommodations are often not relevant or helpful and have limited supported research. It is especially important that accommodations for ELL students must be language related in order to be effective in making assessments more accessible for these students.

One of the major issues for the future is the need to expand the research in the area of accommodations as there is not enough research to judge the effectiveness and validity of many of the existing accommodations for both SD and ELL students. For example, score comparability is highly related to the outcome of accommodated assessment. If provision of accommodation alters the construct being measured, then accommodated assessment outcomes may not be valid and as a result the accommodated and nonaccommodated assessment outcomes cannot be aggregated. Additional research is needed to help schools choose the best accommodations and to ensure that the outcome of accommodated and nonaccommodated assessments can be aggregated. Recent publications reporting results of research on accommodations for ELL and SD students, including the taxonomy of accommodations provided in Rivera and Collum (2006), could help schools make better choices in selecting existing accommodations rather than using a common sense approach in their decisions.

However, it must be noted at this point, that some accommodations may have a limited impact on assessment outcomes and may only be considered a quick fix because they may not be able to systematically address the underlying issue of equitably assessing immigrants and ELL students in providing an assessment in the appropriate language. Other accommodations may help make assessments more accessible – and consequently more valid and fair – for immigrants and ELL students. For example, Levin et al. (2003) found that if bilingual students were

able to take a test in both of their languages, their performance improves because they construct meaning in two languages rather than one. Findings of a study by Levin and Shohamy (2007) indicated that "immigrants, rather than being deficient [in terms of their language resources] have a clear advantage that should be included in an expanded view of the construct of academic language" (p. 19). Obviously, the native language assessment is effective under the condition that the language of instruction and the language of assessment are aligned.

Abedi and Ewers (2013) provided detailed description of the effectiveness and validity of accommodations used for English language learners and students with disabilities and, based on an extensive literature review and experts' advice, created a Research-Based Decision Algorithm based on which decisions can be made on which accommodations can best fit their students with particular academic background backgrounds. The authors created a coding system with the following five categories: (1) "Use," (2) "Use/Low Evidence," (3) "Not Use," (4) "Unsure," (5) "Unsure/Low Evidence Needed," (6) "Unsure/Moderate Evidence Needed," and (7) "Unsure/High Evidence Needed."

Accommodations that are labeled as "Use" are supported by a preponderance of evidence on their effectiveness and validity and they are judged to be relevant and feasible. As an example of accommodations labeled as "Use" Linguistic modification of the assessments can be use. In this approach, linguistic complexities that are judged by content experts to be irrelevant to the focal construct are removed or simplified. Different studies have confirmed that this accommodation does not alter the focal construct.

It is also important to understand how instruction and assessment interact. Students can benefit more when accommodations are provided under both assessment and instruction conditions. This combination provides an opportunity for bilingual and ELL students to become familiar with the accommodations that are used in their assessments.

The concept of academic language is an extremely important consideration when dealing with the assessment of immigrants and English language learners in content-based areas such as math and science. While everyone, particularly immigrants and ELL students, can greatly benefit from assessments with clear language, these students must also be familiar with the language that facilitates content learning, i.e., academic language. For example, as Levin and Shohamy (2007) pointed out, content literacy, rather than language per se, greatly impact students' performance in content-based areas. For example, Levin and Shohamy indicated that, "not only the vocabulary and symbols but also the norms, values, and conventions that are characteristics of the discipline" (p. 18).

Finally, the differential item functioning (DIF) approach may help identify specific items that discriminate against students who are not proficient in the language of assessment. The effectiveness of accommodations can then be examined on test items that exhibited a high level of DIF (C-DIF) (see, for example, Uiterwijk and Vallen 2003).

## Cross-References

- ▶ Assessing Second/Additional Language of Diverse Populations
- ▶ The Common European Framework of Reference (CEFR)
- ▶ Using Portfolios for Assessment/Alternative Assessment

## Related Articles in the Encyclopedia of Language and Education

Sven Sierens, Piet Van Averameat: Bilingual Education in Migrant Languages in Western Europe. In Volume: Bilingual and Multilingual Education.

Laura Sterponi: Language Socialization and Autism. In Volume: Language Socialization.

Kutlay Yagmur: Multilingualism in Immigrant Communities. In Volume: Language Awareness and Multilingualism.

## References

Abedi, J. (2006a). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah: Lawrence Erlbaum Associates.

Abedi, J. (2006b). *Are accommodations used for ELL students valid?* Paper presented at the 2006 annual meeting of the American Educational Research Association in San Francisco.

Abedi, J. (2012). Validity issues in designing accommodations. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing*. London: Routledge/Taylor & Francis Group.

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Accomodations-for-under-represented-students.pdf

Abedi, J., & Hejri, F. (2004). Accommodations in the national assessment of educational progress for students with limited English proficiency. *Applied Measurement in Education, 17*(4), 371–392.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics* (CSE Technical report no. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Abedi, J., Kim-Boscardin, C., & Larson, H. (2000). *Summaries of research on the inclusion of students with disabilities and limited English proficient students in large-scale assessment*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Bennet, R. E., Rock, D. A., & Jirele, T. (1987a). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *Journal of Special Education, 21*, 9–21.

Bennet, R. E., Rock, D. A., & Kaplan, B. A. (1987b). Level, reliability, and speededness of SAT scores for nine handicapped groups. *Special Services in the Schools, 3*(4), 37–54.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille edition. *Journal of Educational Measurement, 26*(1), 67–79.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodation: Effects on multiple-choice reading and math items* (Technical report 31). Minneapolis: University of Minnesota/National Center on Educational Outcomes.

Burns, E. (1998). *Test accommodations for students with disabilities*. Springfield: Charles C. Thomas Publisher.

Chiu, C. W. T., & Pearson, D. (1999, June). *Synthesizing the effects of test accommodations for special education and limited English proficient students*. Paper presented at the National Conference on Large Scale Assessment, Snowbird.

Coleman, P. J. (1990). *Exploring visually handicapped children's understanding of length (math concepts)*. (Doctoral dissertation, The Florida State University, 1990). *Dissertation Abstracts International, 51*(0071).

DiCerbo, K., Stanley, E., Roberts, M., & Blanchard, J. (2001, April). *Attention and standardized reading test performance: Implications for accommodation*. Paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC.

Elliott, S., Kratochwill, T., & McKevitt, B. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology, 31*(1), 3–24.

Francis, D. J., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners. Research–based recommendations for the use of accommodations in large-scale assessments*. Houston: Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston for the Center on Instruction.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments about test accommodations with objective data sources. *School Psychology Review, 29*(1), 65–85.

Gallina, N. B. (1989). Tourette's syndrome children: Significant achievement and social behavior variables (Tourette's syndrome, attention deficit hyperactivity disorder) (Doctoral dissertation, City University of New York, 1989). *Dissertation Abstracts International, 50*(0046).

Gándara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. *Education Policy Analysis Archives, 11*(36). Retrieved 14 June 2006 from http://epaa.asu.edu/epaa/v11n36/

Hafner, A. L. (2000, April). *Evaluating the impact of test accommodations on test scores of LEP students & non-LEP students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English-language learners in U.S. Schools*. Washington, DC: National Academy Press.

Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take English language learners to attain proficiency?* Santa Barbara: University of California/Linguistic Minority Research Institute.

Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1999). *Handwritten vs. word processed statewide compositions: Do judges rate them differently?* Eugene: University of Oregon, BRT.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. (2009). Accommodations for English language learners taking large-scale assessments: A Meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services, 2000–2001 summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring math – Not reading – On a math assessment: A language accommodations study of English language learners and other special*

*populations*. Presented at the annual meeting of the American Educational Research Association, New Orleans.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (Technical report no. 431). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Levin, T., & Shohamy, E. (2007). The role of academic language in understanding the mathematics achievements of immigrant students in Israel. In C. S. Sunal & K. Mutua (Eds.), *The enterprise of education: Research on education in Africa, the Caribbean, and the Middle East* (pp. 313–336). Tuscaloosa: Info Age Publishing.

Levin, T., Shohamy, E., & Spolsky, D. (2003). *Academic achievements of immigrant students: Findings and recommendations for decision makers*. Ministry of Education/Department of the Cheif Scientist [in Hebrew].

Li, & Suen. (2012). *Theoretical Economics 7*, 357–393

Mac Arthur, C. A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *Journal of Special Education, 21*(3), 22–42.

Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students*. Paper presented at the Council of Chief State School Officers 32nd annual national conference on Large-Scale Assessment, Palm Desert.

Marquart, A. (2000). *The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students of various skill levels*. Paper presented at the annual meeting of the Council of Chief State School Officers, Snowbird.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES publication no. 2000–473). Washington, DC: National Center for Education Statistics.

Meloy, L. L., Deville, C., & Frisbie, C. (2000). *The effect of a reading accommodation on standardized test scores of learning disabled and non-learning disabled students*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Mick, L. B. (1989). Measurement effects of modifications in minimum competency test formats for exceptional students. *Measurement and Evaluation in Counseling and Development, 22*, 31–36.

Moore, K. A., & Redd, Z. (2002). *Children in poverty: Trends, consequences, and policy opinion*. Washington, DC: Child Trends Research Brief.

Munger, G. F., & Loyd, B. H. (1991). Effect of speededness on test performance of handicapped and non-handicapped examinees. *The Journal of Educational Research, 85*(1), 53–57.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Perez, J. V. (1980). Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation (Doctoral dissertation, University of South Florida, 1980). *Dissertation Abstracts International, 41*(0206).

Phillips, S. E. (1994). High stakes testing accommodations: Validity vs. disabled rights. *Applied Measurement in Education, 7*(2), 93–120.

Rivera, C. (2003). *State assessment policies for English language learners*. Presented at the 2003 Large-Scale Assessment conference.

Rivera, C., & Collum, E. (2006). Including and accounting for English language learners in state assessment systems. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective*. Mahwah: Lawrence Erlbaum Associates.

Rivera, C., & Stansfield, C. W. (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998–1999*. Arlington: The George Washington University/Center for Equity and Excellence in Education.

Rogers, W. T. (1983). Use of separate answer sheets with hearing impaired and deaf school age students. *B. C. Journal of Special Education, 7*(1), 63–72.

Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives, 7*, 1–47.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives, 5*(3), 1–20.

Russell, M., & Plati, T. (2001). *Effects of computer versus paper administration of a state-mandated writing assessment*. TCRecord.org. Retrieved 23 Jan 2001, from the World Wide Web http://www.tcrecord.org/PrintContent.asp?ContentID=10709

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodations on test performance: A review of the literature*. Center for Educational Assessment (Research report no. 485). Amherst: School of Education, University of Massachusetts Amherst.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13.

Sullivan, P. M. (1982). Administration modifications on the WISC-R performance scale with different categories of deaf children. *American Annals of the Deaf, 127*(6), 780–788.

Thurlow, M. L. (2001, April). *The effects of a simplified-English dictionary accommodation for LEP students who are not literate in their first language*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis report 41), University of Minnesota, National Center on Educational Outcomes, Minneapolis, Retrieved 13 July 2002, from the Website http://education.umn.edu/NCEO/OnlinePubs/Synthesis41.html

Thurlow, M., & Liu, K. (2001). *State and district assessments as an avenue to equity and excellence for English language learners with disabilities* (LEP projects report 2). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000) *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis report 33). Minneapolis: University of Minnesota/National Center on Educational Outcomes. Retrieved 9 Oct 2006 from the Website http://education.umn.edu/NCEO/OnlinePubs/Synthesis33.html

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children, 64*(4), 439–450.

Tippets, E., & Michaels, H. (1997). *Factor structure invariance of accommodated and non-accommodated performance assessments*. Paper presented at the National Council on Measurement in Education annual meeting, Chicago.

Uiterwijk, H., & Vallen, T. (2003). Linguistics sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 2*, 211–234.

Varnhagen, S., & Gerber, M. M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly, 7*, 266–270.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students* (Minnesota report 34). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Watkins, M. W., & Kush, J. C. (1988). Assessment of academic skills of learning disabled students with classroom microcomputers. *School Psychology Review, 17*(1), 81–88.

Willner, L. S., Rivera, C., & Acosta, B. D. (2009). Ensuring accommodations used in content assessments are responsive to English-language learners. *Reading Teacher, 62*(8), 696–698.

Wolf, M., Kim, J., & Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Applied Measurement in Education, 25*(4), 9347–9373.

Wright, N., & Wendler, C. (1994). *Establishing timing limits for the new SAT for students with disabilities*. Paper presented at the annual meeting of the National Council on Measurement in Education (New Orleans, 4–8 Apr 1994). ERIC ID# ED375543.

# Assessing the Language of Young Learners

Alison L. Bailey

**Abstract**

Developmental realities of young language learners (approximately ages 3–11 years) pose unique and challenging considerations for oral and written language assessment. This chapter first addresses construct definitions important for establishing a common understanding of language testing with young children. Both type (e.g., summative, formative) and purpose (e.g., accountability, diagnostic) of language testing commonly used in preschool and elementary (primary) school contexts are discussed. Developmental and cultural factors that need to be taken into account in assessing young learners are addressed. I conclude with work in progress and future directions for language test development involving empirically derived trajectories of language development, accounting for the intersection of language and content learning, the professional development of teachers around language learning and assessment, and further innovations with technology.

**Keywords**

Age-appropriate assessment • Developmental trajectories • Proficiency • Young learners

## Contents



A.L. Bailey (✉)
Human Development and Psychology Division, Department of Education, University of California, Los Angeles, CA, USA
e-mail: abailey@gseis.ucla.edu; bruinprofab@gmail.com

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_22

## Introduction

The field of language testing has recently seen new volumes and chapters dedicated to discussion of assessment with young language learners (e.g., Bailey et al. 2014; Nikolov 2016), compendiums characterizing and evaluating available language assessments for young learners (Barrueco et al. 2012; Jang 2014), and a burgeoning research agenda that as included studies of student self-assessment, the intersection of language and academic content learning, and further exploration of technology-assisted assessment with children. In part, this is likely a consequence of the increased communicative demands placed on school-age language learners. Young language learners not only encounter the assessment of their language development but also assessment of other learning and knowledge through the language(s) they are still acquiring. For example, students are now being expected to display academic content knowledge through oral and written explanations and argumentation both in daily classroom-based tasks and in summative assessments that are tied to new academic content standards (e.g., Bailey and Heritage 2014).

The shift in expectations for language demands has been led in the USA primarily by the Common Core State Standards Initiative of the National Governors Association Center for Best Practices, Council of Chief State School Officers (CCSSO 2010), and Next Generation Science Standards (NGSS Lead States 2013) and by similar accountability initiatives in other countries, for example, in the UK, new tests aligned to the National Primary Curriculum are expected in 2016 (Department for Education 2016)

Such explicit expectations for language competencies make a chapter devoted to the language assessment of the very youngest language learners more critical than ever. More than a decade ago, Inbar et al. (2005) called for "specific age appropriate and language level considerations" (p. 3.) in order to address the differences between testing the language learning population at large and testing young language learners. Much of what has been examined elsewhere in this volume is reconsidered

in this chapter from the points of view of those who must create valid (i.e., fair and effective) tests for assessing the language of young learners and those who must administer and interpret them. These viewpoints require familiarity with testing purposes and an understanding of developmental and cultural issues as they impact the design and use of language assessments with young children.

While not exclusively the case, this chapter deals predominantly with tests of students' English language development (ELD) or proficiency (ELP). This is a reflection of both the increasing number of young children learning English in various contexts around the world (Graddol 2006) and the fact that much research has been conducted on the assessment of English.

The chapter is organized around five main sections: First, I provide construct definitions that will prove important for establishing a common understanding of testing issues with young children, starting with a definition of the term "young learner" itself. Second, I review the types (e.g., summative, formative) and purposes (e.g., accountability, diagnostic) of language testing in preschool and elementary (primary) school contexts. Third, I address the developmental child level concerns that need to be taken into account in assessing this population of test takers, including a review of general guidelines and best practices for assessing young children. In the fourth section, I consider culture as an additional contextual factor that, while possibly impacting all language testing situations, may have particular significance for the testing of young children. Finally, in the fifth section, I conclude with updates on how the field has progressed over the past decade, current works in progress, and future directions for test development.

## Construct Definitions

Many key constructs already encountered in other chapters will need special definition in the context of assessing young language learners.

### Defining Young Language Learner

I start with the most crucial of all definitions for this chapter, that of the *young language learner*. Defining young language learner is complicated by the range of language learning experiences, the range of ages to be covered by the qualifier "young," and the fact that in different parts of the world, different school systems introduce students to second language and foreign language instruction at different points in their school careers. In Europe, young learner is often applied to students in only the very earliest school years (ages 5–7) or before. In the USA, where the introduction of foreign language teaching often does not take place until the secondary grades, the notion of a "young learner" can span the entire preschool and elementary years (ages 3–11). Obviously for second language learners, the

onset of a second language can start before the start of formal schooling or at any time during the primary school years for those who emigrate as school-age language learners.[1] Much of the focus of this chapter, however, will be on young learners from preschool through the earliest elementary years.

## Defining the Language Learning Context

Turning next to language construct definitions, prominent among these are English as a second or additional language (ESL or EAL), bilingualism, and due to the demand for English in non-English- speaking countries around the world, English as a foreign language (EFL).[2] Second and foreign languages other than English will also be pertinent to a broader discussion of young language learners everywhere. Language assessment for young monolingual speakers is primarily confined to the literate uses of a language (e.g., reading and writing), with the exception of instances when a language disability is suspected or has been diagnosed for intervention and monitored for improvement.[3] There are few assessments of oral language proficiency in a first (and often only) language, and yet the increased language demands placed on students in schooling contexts will affect *all* students not just those learning an additional language.

Second language acquisition (SLA) such as ESL is made more complex in the young learner context by the existence of bilingual first language acquisition (BFLA) (De Houwer 1998), in which children may be acquiring two languages, each as a native language. As they enter formal schooling environments, including preschool, these children may become literate in only one of their two languages if the schooling system favors one language over the other or if parents do not opt to enroll their children in dual language programming. The language learning experiences of young children may also be characterized by immersion in a second language they are yet to acquire. In Canada, for example, children have the opportunity to learn English and French (and other desired languages) in this environment from an early age (see Bailey and Osipova 2016 for review of educational options with young language learners).

EFL (and other foreign language acquisitions) characterizes learners who acquire a language after their native language has already been acquired, but do so outside an

[1]Young learners of English are variably referred to as English learners (ELs), English language learners (ELLs), English as a second (ESL) or additional language (EAL) students, students with non-English-speaking backgrounds (NESB), or, most recently, emergent bilinguals or dual language learners (DLLs) to reflect that many young students are acquiring English and a home language (García 2009). Throughout this chapter, I simply use young language learner.

[2]ELD will be used throughout this chapter to refer to the process of English acquisition regardless of whether it is being acquired as a second or foreign language.

[3]Review of such clinical assessments is outside the scope of this chapter (see Conti-Ramsden and Durkin 2012 and Dockrell and Marshall 2015 for recent reviews of language assessment with children with language learning disabilities).

English (or other target L2) environment. For the very youngest preliterate language learners, this may mean learning a foreign language without the aid of the print medium that is available to older children and adult learners. Older learners can garner literacy abilities in their L1 to augment their learning of oral English, as well as transfer print skills in their L1 to reading and writing in English. The latter is particularly enhanced if their L1 shares the same orthography and possibly even cognate words with English.

## Defining Language Varieties

This chapter adopts a broad definition of language including all four modalities of listening, speaking, reading, and writing and, where relevant, further denotes sub-skills such as phonological awareness and pronunciation. Additional construct definitions that need to be taken into account in the assessment of young learners include the social and academic language constructs (Cazden 2001; Chamot 2005). While the distinction between the language used in a scholastic environment and the language used in everyday (out-of-school) contexts may not be as great during the early years of schooling as it is once children begin to take discipline-specific classes (e.g., history, algebra), the distinction arguably still exists. With increasing preschool enrollment worldwide, more young children have been affected by ties between opportunities for preschool language development and later academic outcomes. Working in preschool settings in Europe, Michel and Kuiken (2014) have found that preschool environments place unique demands on the language of young learners and consequently require appropriate ways to assess the language development of the very youngest of students.

The existence of an academic language construct is not without controversy, however, especially in what constitutes fair assessment of the obvious scholastic uses of language at this young age – emergent reading and writing. *Should the reading and writing skills in English of young learners be assessed differently from those of native English students who are also just beginning to learn to read and write?* If young English learners are already literate in their L1, there are implications for how we assess their literacy in English. Environmental print (i.e., sight words) from the content areas such as science, mathematics, and history may make appropriate content for assessing the literacy abilities of young school-age learners. However, there may be no positive transfer for literacy skills from children's L1 to their L2 if the orthographies of the two languages do not match (Bialystok et al. 2005), although more recently Gottardo et al. (2006) report significant correlations in the phonological processing of young language learners whose languages do not share orthographic systems (e.g., Chinese and English). Reading and writing modalities may, however, still be problematic in other ways when operationalized for testing their development in young children. For example, reading and writing are frequently tested orally which requires children to *listen* to directions not simply demonstrate their literacy abilities. Conflating these skills may result in ambiguous information for teachers wishing to effectively target their instruction. Finally, the

academic language construct may not be imperative for acquiring and displaying learning in content areas when children can effectively convey their mathematics, science, history learning, etc., using all linguistic resources at their disposal including use of L1, as well as everyday and nonstandard varieties of L2 (e.g., Faltis 2015).

## Types and Purposes of Assessment

As with assessments developed for use with older children and adults, there is a range of purposes for language assessment with young learners. Due to the maturational constraints and the need for developmentally relevant measures, we witness far greater variety in the purpose and use of informal assessment in this young population.

### High-Stakes Assessment

With standardized assessments, the content is a sampling of all that a student may have been taught in a given period. These assessments are summative of knowledge gain and are often considered "high stakes" for the student (e.g., a deciding factor in being reclassified as a fluent English speaker for instructional placement) or "high stakes" for those who educate them (e.g., evaluation of teacher or school performance). Also considered "high stakes" but not summative are assessments designed to screen a student's abilities for weaknesses that need immediate amelioration or flagged for possible future attention. Such screening purposes can also be considered "high stakes" for both the individual and the schooling system. An individual needs to be accurately identified for further instruction or services if these are necessary to their development. These are the cases when the schooling system also needs accurate information; providing services to individual students who are falsely identified as in need of services will not be cost effective, and those who are falsely identified as sufficiently able when they are not may require more costly remediation at a later point in time (Vellutino and Scanlon 2001). Technical quality of a test in terms of validity and reliability and the integrity of young learner language assessment systems as a whole are of course major considerations when the stakes for testing young students are high (McKay 2005; Bailey and Carroll 2015).

### Assessment for Learning

Assessment for instructional or diagnostic purposes can take the form of standardized summative assessments or classroom-based formative assessments. Standardized assessments will offer the language teacher information about a sample of items across a variety of domains to measure general language proficiency or within a single domain of language such as vocabulary or syntax and how well a student is

doing on these skills relative to either standards (i.e., criterion referenced) (e.g., the TOEFL Primary developed by Educational Testing Service for children aged 8 and older is mapped to the Common European Framework of Reference (CEFR, Council of Europe 2001)) or relative to other students his or her age, grade, or level of overall language proficiency (i.e., norm referenced) (e.g., the *pre*LAS developed by CTB McGraw-Hill for assessing 4–6-year-olds in both English and Spanish as either an L1 or L2). The information gained can be used to monitor annual progress or to categorize students for educational purposes, for example, to literacy instruction in a child's dominant language. However, the information from such standardized assessments is likely to be neither sufficiently refined nor contain a critical number of like items to effectively target specific subskills. Educators must guard against using information from tests designed for one purpose (e.g., annual growth in general language proficiency) with another purpose in mind (e.g., next-steps instructional decisions) (National Educational Goals Panel [NEGP] 1998). Alternative or formative assessment is, however, designed to closely guide student learning as Wiliam (2006) explains:

> What makes an assessment formative, therefore, is not the length of the feedback loop, nor where it takes place, nor who carries it out, nor even who responds. The crucial feature is that evidence is evoked, interpreted in terms of learning needs, and used to make adjustments to better meet those learning needs. (p. 285)

Assessment for learning, such as formative assessment, is especially pertinent in the case of young learners still acquiring a new language. Formative approaches to assessment can capture a broad array of relevant language information for teachers that is closely tied to the young learners' instructional needs (Davidson and Lynch 2002; Frey and Fisher 2003). Formative assessment can be conducted by teachers either informally while "on the run" as part of ongoing instruction, or it can be formal, that is, planned in advance to address certain aspects of student language knowledge (e.g., McKay 2006). A central focus of formative assessment is teacher feedback to students, as well as a focus on student monitoring of their own language learning through self-assessment (Bailey and Heritage 2008).

Formative assessment may also include extra-child characteristics such as the classroom environment, parental involvement, home literacy habits, etc., and take many different forms (see Tsagari 2004 for a brief overview of the nomenclature and strengths and weakness of alternative assessments in the language assessment context). The use of informal observations, for example, allows for a range of skills (e.g., peer-to-peer oral discourse) not always amenable to more formal or standardized assessment environments. Observations can also be made formally and used to evaluate the quality of the language environment of a classroom rather than individual students (e.g., the Sheltered Instruction Observation Protocol, SIOP, Echevarria et al. 2004). The use of progress maps on a developmental continuum in order to estimate a student's growth over time (Masters and Forester 1996) and the use of portfolios to create individual profiles of language learning progress and achievement (e.g., Butler and Stevens 1997; CEFR, Council of Europe 2011;

Puckett and Black 2000) are alternative methods well suited to documenting the language of young learners and facilitating teachers' decision making for further learning. Such approaches can even be adopted by students themselves. For example, the "language passport" supported by the Council of Europe's European Language Portfolio initiative (2011) is used by students to directly rate their own language proficiency, although see Hasselgreen 2005 for a critique of the CEFR with younger language learners to which the European Language Portfolio is mapped.

## Developmental Considerations

Motivation for this chapter comes primarily from the recognition that there are developmental and contextual factors that must be taken into account with the assessment of young language learners (e.g., Inbar et al. 2005; McKay 2006; Rea-Dickins and Rixon 1997). As in the USA, initiatives in Australia, Canada, and the UK have placed increasing emphasis on school systems to be held accountable for monitoring progress in the language development of young students, particularly young immigrant or language minority students (Indigenous and nonIndigenous) (e.g., McKay 2005, 2006; Silburn et al. 2011). There has also been an increase in young children studying English as a foreign language in non-English-speaking countries. Graddol (2006) reports that:

> The age at which children start learning English has been lowering across the world. English has moved from the traditional 'foreign languages' slot in lower secondary school to primary school – even pre-school. The trend has gathered momentum only very recently and the intention is often to create a bilingual population. (Graddol 2006, p. 88)

An interesting prediction stemming from this situation is that in the future there will be only "young" learners of English as older members of societies will have acquired English earlier in life. Consequently, it is appropriate that learners in this young age range receive emphasis in future assessment development and research efforts.

In a review of research on the assessment of school-age language learners conducted in various parts of the world, McKay concludes that young learner assessment deserves to be established as a highly expert field of endeavor requiring, for example, knowledge of the social and cognitive development of young learners, knowledge of second language literacy development, and understanding of assessment principles and practices (McKay 2005, p. 256). Beginning with McKay's assertion that the field develop an understanding of assessment principles and practices, three main areas of test design with young children require special consideration: (1) format (whether individual, small group, or whole class), (2) choice of item and task types, and (3) choice of contextualized, age-appropriate stimuli (Inbar et al. 2005). Explicitly identifying these three areas raises specific challenges for test development practices with young children. In each

case, design decisions must take the learning context into account to establish a match between instructional environment and assessment.

## Test Format

The language modality and age of the test taker will certainly dictate the appropriate format in which to assess young learners. Individual assessment will be necessary for coverage of many of the skills in the speaking and listening modalities. However, in the preschool setting, many classroom teachers also call upon children to respond in unison (e.g., sing-alongs, calling out keywords as a chorus, and providing en masse actions/enactments to stories and poems, Tabors 2008). A child's ability to both comprehend and participate in such group activities should be at least one focus of assessment with the youngest language learners in this early instructional context.

Assessment of early literacy may need to be carried out in individual or in small group contexts because test takers cannot be relied upon to be sufficiently proficient to read directions for responding to print items or tasks nor to maintain their attention in group settings. No matter the format, limiting the duration of the test to avoid testing fatigue will be of far greater concern with this young population than with older test takers.

## Test Item and Task Types

Choice of item and task types will need to correspond to the cognitive processing capabilities and degree of task familiarity of young learners. For example, Weir (2005) provides a language test item that requires making meaning from a bar chart. This type of task presents statistical information in a way that primary school children will encounter in graphics during mathematics, science, or social studies lessons. An item type that requires responding to a series of questions based on information extracted from a graphic would be appropriate once children have, as part of the school curriculum, received explicit instruction in "reading" graphic displays of information, otherwise the item type would be unfamiliar, and the task demands too great for the young learner; simply put, the demands of the assessment should match the demands of the curriculum. Other cognitive developments that will restrict the range of tasks include attention span and memory. For example, multistep items that require sequential manipulations of information or lengthy passages of text followed by comprehension questions may be outside the cognitive capacity of the youngest language learners.

To lessen the negative impact of processing demands and to capitalize on the degree of assistance learners may require from more expert others (Vygotsky 1978), assessments can award partial credit based on the verbal scaffolding necessary to elicit a response from young ESL test takers. This strategy allows for diagnostic information to be generated. The differing levels of response reveal how much

knowledge a child has and how much they still need to learn to succeed without assistance.

### Task Content

The content of the tasks needs to be relevant to the young learner in terms of cognitive demands and cultural specificity (culture is addressed further in the later section). The younger the learner, the more contextualized the items will need to be in order for the test taker to make meaning of them. That is, items will need to be topically appropriate for the target age of the test taker, and the ability to answer the items should not require knowledge of information not already provided in the tasks or test items. Cognitive developments impacting these considerations include an awareness of testing procedures or the "test genre" (i.e., cooperation in attempting to answer all items and providing adequate constructed responses), as well as an understanding of an opportunity to use decontextualized language – that is, responses will be sufficiently explanatory for the absent test grader to make meaning of them. For the youngest learners, a number of easy "warm-up" items can be used to familiarize the child to the tester and testing procedures and should not be scored. Manipulatives (i.e., toy farm animals, dolls) can be incorporated in both item questions and response formats. According to research, young children are more successful on both production and comprehension tasks if the tasks use objects rather than pictures (e.g., Cocking and McHale 1981; Serna 1989 cited in Beck 1994); objects help contextualize the task in the cognitively less demanding "here and now."

Choice of age-appropriate content in test construction is made more complex with young learners than in other testing target groups because language development is concurrent with developments in other areas (e.g., scholastic, cognitive, and social developments). Because a child may begin learning a second or foreign language at any point in their early school years, the development of the language can be asynchronous with developments in other areas. The beginner status of young learners in the later elementary grades makes choosing content difficult (i.e., restrictions on availability of age-appropriate topics from which to select beginning level vocabulary and simple discourse contexts). This is also the situation if a test is to span an age range rather than be targeted at individual grades or ages.

## Impact of Child Development on Test Interpretation

Cognitive and social developments not only impact test design but also the manner in which tests are administered and interpreted. Assumptions upon which validity arguments are made with standardized assessments (e.g., Davidson and Lynch 2002; Weir 2005) are often compromised when administering such tests with young children. For example, the assumption of uniformity of the testing experience for all the test takers is not met with young children whose attention abilities and

familiarity with test taking can vary tremendously (Powell and Sigel 1991). Moreover, what is considered "typical" for this young age range also varies. This raises the issue of whether using certain types of assessment with very young children is desirable. If the purpose of assessment is accountability of the program, then group level and classroom-related indicators (e.g., amount of student engagement, ESL experience of teaching staff) may be most appropriate. If the purpose is diagnostic, then information on individual students may be preferred. However, caution is required because of the compromises outlined earlier. Interpretations from formative assessment approaches rather than from administration of standardized assessments may be more meaningful.

## Guidelines for Assessing Young Children

While not specifically targeting the assessment of language, there are several general test administration guidelines for use with young learners. For the very youngest learners in the USA, the *Principles and Recommendations for Early Childhood Assessments* assembled by the NEGP (1998) still hold. These include but are not limited to the following four guidelines: (1) assessments should be tailored to a specific purpose and should be reliable, valid, and fair for that purpose, (2) policies should be designed recognizing that reliability and validity of assessments increase with children's age, (3) assessments should be age appropriate in both content and the method of data collection, and (4) parents should be a valued source of assessment information, as well as an audience for assessment results. Specific guidelines for assessment practices with young English learners have also been published by the National Association of the Education of Young Children (NAEYC 2009).

As the basis for all assessment (clinical, scholastic, and linguistic) in K-12 education in the USA, the *Standards for Educational and Psychological Testing*, published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2014), includes comprehensive guidelines for the testing of individuals of diverse linguistic backgrounds, and researchers working with language minority students have additionally made contributions to guide fair and valid assessment practices of both the language and other knowledge areas of language learners (see also Sireci and Faulkner-Bond for review 2015).

Suggested practices include use of test accommodations. Accommodations such as extra time and dictionaries are thought to vitiate the interpretation of test results obtained with ELL students as long as these have been empirically proven not to alter the language construct to be measured (e.g., reading comprehension). Increasingly technology can play a role in test accommodations such as the use of computer-administrated bilingual dictionaries or glossaries (e.g., Abedi 2014). If the construct is not language ability itself but the test uses language as the medium (e.g., an assessment of mathematics requiring both reading and writing skills), then accommodation options include not only extra time and bilingual dictionaries/glossaries but also the option to test in a student's native language if this matches the language

of instruction (Pennock-Roman and Rivera 2011) (for a general discussion, see also Abedi, chapter "▶ Utilizing Accommodations in Assessment," Vol. 7). However, interpretation of accommodated results as valid indicators of academic content area knowledge has real consequences for students; we should still be cautious in our interpretations because as Davidson (1994) points out, norming studies of such academic achievement assessments have often not included learners who reflect the full range of language proficiencies found in schools.

Other impacts on administration and interpretation of language assessments include the training needs of teachers who often must administer assessments to school-age language learners. General education teachers have been found to have little training in language development and assessment (e.g., Téllez and Mosqueda 2015). Scoring and reporting the test performance of young learners also proffer challenges to teachers and test developers alike. Scoring concerns include the degree of teacher variance in what is considered an acceptable answer. For example, children's immature articulatory abilities or their productions influenced by L1 may make responses difficult to decipher and thus score reliably.

Multiple sources of evidence should be used to increase the validity of inferences about student language performances (e.g., Conti-Ramsden and Durkin 2012; Dockrell and Marshall 2015). Employing multiple measures helps prevent overreliance on any one assessment that may yield a biased view of performance due to cognitive or social development constraints. This does not however entail administering large batteries of standardized assessment that could quickly lead to test fatigue in young children. Rather, studies of expert teachers suggest that they use their knowledge of teaching and learning to create an ongoing cyclic process of teaching and assessment involving a repertoire of both formal and informal assessments (Rea-Dickins 2001). For example, evidence of language proficiency can come from combining a student's performance on formal assessments and informal quizzes and from teacher observation during class time (e.g., Frey and Fisher 2003), as well as utilizing self- and peer assessment.

Finally, reporting the results of a test performance to young children also needs to be carefully considered and made age appropriate to avoid issues of demotivation or threats to a child's self-esteem. However, reporting results to children and reporting results to teachers and parents need not be the same process, and teachers will need item level or subskill level information from assessments in order to make effective instructional modifications.

## Impact of Culture on Administration and Interpretation

Culture impacts the fair and valid testing of young children's language abilities when there is a mismatch between home practices in communication and those practices commonly used for assessment. For example, Peña and Quinn (1997) report that Latina and African-American mothers typically do not label objects in their children's environment (as is the case for most vocabulary assessment), but rather engage in games that more often require descriptions. Thematic content of an

assessment also needs to be compatible with children's home culture (at least culturally appropriate for the majority of learners taking the test, if known). Alternatively, assessors have successfully administered dynamic assessments using a test-teach-test design with preschool children to reduce bias from lack of cultural familiarity with vocabulary (Peña et al. 2001). In addition, many children come from backgrounds where they might be expected to learn from observation rather than overt participation and to demonstrate their comprehension nonverbally (e.g., Beck 1994; Scollon and Scollon 1981). Collectively, this research should impact test development design, encouraging more development of dynamic assessments, or in the case of listening comprehension, creation of items that do not rely exclusively on verbal responses to signal accurate comprehension. Furthermore, early childhood education agencies, such as the Head Start in the USA, recommend caution with the interpretation of language assessments with young children noting the need to assess and take into account all the languages a young learner knows during educational decision making (Office of Head Start 2010) (for discussion, see Mahon et al. 2003).

## Work in Progress and Future Directions

The twenty-first century began with a new era of educational accountability impacting young language learners in terms of the language demands now placed on them in schooling contexts. The field has shifted from recognizing that assessing young learners entails the assessment of their language development and the assessment of their academic content learning through language to also recognizing the need to take account of the language practices of the academic content areas on those very language assessments themselves. This has led to recent comprehensive language test development efforts in many parts of the world. In the USA, "next-generation" ELD assessment is under way under federal government initiatives to align ELD assessment with the academic content standards. Add to this mandate the anticipated expansion of publicly funded education to young, preschool-age children, many of whom are the children of immigrants from non-English-speaking countries. In Australia, add to this the increased focus on the language learning needs, indeed rights, of Indigenous students (Silburn et al 2011). In Europe, add in the expansion of the European Union with the increased mobility this brings, as well as, most recently, asylum seekers from the Middle East and North Africa, and collectively large numbers of families with young children are settling in areas of Europe where they do not speak the dominant language. Much research and test development has still to be done to improve assessment of the wide range in language demands now facing young learners.

Recommendations made in the 2008 version of this chapter of the encyclopedia to meet the language assessment needs of young language learners were organized around three aspects of the mission statement of the Committee on Early Childhood Pedagogy (National Research Council 2001). Reviewing those aspects now (technical quality of assessments, teacher professional development around language assessment, and integration of technology) reveals to what extent advancements

have been made nearly a decade on and which areas still need the attention of researchers and educators.

In terms of advancing the technical quality of assessments and how they are used to support the learning of young learners, much has been achieved in articulating construct definitions of necessary language knowledge and skills. The call for the revision or creation of ELD standards for the preschool through school-age levels has not only been met by several individual US states and consortia, for example, but CCSSO created the *Framework for English Language Proficiency Development Standards* (CCSSO 2012) to guide such revisions based on a synthesis of research around language and content learning (e.g., Lee et al. 2013) that has led to the identification of key *language practices or performances* found to be common across the new language arts and mathematics and science standards (CCSSO 2010; NGSS 2013). These practices and performances are a "combination of communicative acts (e.g., saying, writing, doing, and being) used in the transmission of ideas, concepts, and information in a socially mediated context" (p. 2) that include, among others, for language arts the support of "analyses of a range of grade level complex texts with evidence," for mathematics "construct viable arguments and critique the reasoning of others," and for science the necessary language to "plan and carry out investigations" and "engage in argument from evidence." Continued research at the intersection of content knowledge and language will no doubt help to refine the construct for future assessment development with this age range. Uccelli and colleagues (2014) are focusing on language that is common across various disciplines at the upper elementary level and how best to assess this construct, whereas others are looking at the intersection of content knowledge and ELD in the preschool context (e.g., the Literacy and Academic Success for English Learners through Science, or LASErS program of the Education Development Center) which could aid us in understanding how content knowledge itself shapes language use.

Accommodations research has also continued apace with, as mentioned, new meta-analyses providing details about the efficacy of accommodation use under different conditions (e.g., Pennock-Roman and Rivera 2011). This nuanced information has informed new principled accommodation guidelines or algorithms for use with school-age students still acquiring the language in which their academic content knowledge will be assessed (Abedi and Ewers 2013; for a general discussion, see also Abedi, chapter "► Utilizing Accommodations in Assessment," Vol. 7).

In other areas dealing with the technical quality of language assessments with young learners, work is still in progress. Despite our calling for developmental trajectories for language acquisition, the field still knows little about the progression of language in young school-age language learners (Hoff 2013). The characterization of language development is paramount in the creation of effective language proficiency assessments. Work under way by the Dynamic Language Learning Progressions (DLLP) project (Bailey and Heritage 2014) addresses the lack of empirically derived trajectories of language development by sampling oral and

written language practices as outlined by the CCSSO *framework* (e.g., explanations of mathematics task procedures) with students aged 5–12 who have varying proficiencies of ELD. This kind of evidence-based approach to creating language progressions needs to be extended to additional language practices (e.g., argumentation), a wider range of academic content areas (e.g., science, history), and of course to students across all grades.

Language learning progressions hold promise not only for informing development of standardized assessments but also for the area of formative assessment. While traditional notions of validity and reliability cannot be easily applied to establishing the technical quality of formative assessment approaches, criteria for establishing the effectiveness of formative assessment in the classroom can be created, discussed, tried out, and refined. Indeed, recently Heritage (2013) has highlighted the immediacy or proximate timing of evidence of student learning as a key facet of what makes formative assessment valid, along with the need for formative approaches to assessment to yield insights into students' current learning that are sufficiently tractable to be useful in instruction. Also within the area of formative assessment, there is accumulating evidence from a program of research on self-assessment that this population of language learners is not too young to benefit from the self-reflection entailed by self-assessment practices. Butler and Lee (2010), for example, found that 11–12 -year-old students in an EFL context were able to improve their English performances and increase their confidence in learning English with regular use of self-assessment in a classroom context. And in-progress work by Pitsoulakis and Bailey (2016) is revealing that children as young as 7 years are able to self-assess with the appropriate scaffolds to notice features of their own language productions.

Research on teacher professional development remains a key area in language assessment with young learners. Many of the guidance documents cited thus far have educators in mind for special caveats to the administration and interpretation of standardized assessments with young children (e.g., NAEYC 2009). Developing teacher capacity around practices that generate evidence of student learning and lead to accurate interpretations remain important topics for language assessment research specifically (Téllez and Mosqueda 2015; Michel and Kuiken 2014) and for formative assessment research more broadly (Wiliam 2006; Heritage 2013). While some early research has shown expert teachers to effectively use assessment for learning (e.g., Rea-Dickins 2001), more research is needed in the area of professional development to answer the question: *How do teachers effectively implement and use a wide repertoire of assessments for a variety of summative and formative purposes?*

Within the past decade, technology has changed the landscape of language assessment, and this is as true for the youngest learners we have considered here as it is with the assessment of older children and adults. The move to computer-based assessment has been made by the standardized assessments already mentioned (e.g., TOEFL Primary Speaking section), as well as by other newly released assessments

such as the Test of English Language Learning (TELL) progress monitoring application from Pearson and the revised ACCESS for ELLs, the annual summative assessment used by the WIDA ELP assessment consortium in more than 35 states.

Electronic devices are also readily available for continuous digital documentation of student progress (Pellerin 2012) with the possibility for even very young children to use the same tablet devices deftly for assessment purposes. Technology is especially suited to the assessment of young children (see also Chapelle and Voss, chapter "► Utilizing Technology in Language Assessment," Vol. 7); the graphic capabilities that technology offers can also provide a child-friendly context for assessment, with testing made enjoyable for young test takers by mimicking familiar games or cartoons.

Technology has solved a key issue in formative assessment – that of data capture, storage, and management. Data management systems can help make formative assessment practices more effective by systematizing the information that teachers may record formally or "on the run." Language corpora can now also be accessed to provide audiovisual and transcript data that provide teachers who lack familiarity with students from diverse language backgrounds with ways to more accurately compare and evaluate their young language learners. The DLLP project in progress has this as an explicit goal of the project (Bailey and Heritage 2014; Bailey et al. 2016). Authentic language use found in linguistic corpora can also be used to guide test item writers in the production of stimuli texts and test questions However, in standardized assessment development contexts with young learners as it has in test development with adult learners (Frantz et al. 2014).

We can assuredly claim that the assessment of young language learners has made large strides toward evolving into what McKay called "a highly expert field of endeavor." The field has the attention of many national governments due to the increased accountability placed on the role of language in the educational outcomes of all young learners but especially those speaking languages other than English or their society's dominant language. This situation has posed challenges on how best to design assessments that are fair and valid with young children, illuminated gaps in our understanding of the intersection of language and academic content learning, required that we continue to learn how to build the capacity of teachers to both summatively and formatively assess their students' language learning, and led us to continue to leverage technology to meet these myriad objectives. the second decade of the twenty-first century is first and foremost an exciting time to be working with young language learners and their dynamic assessment needs.

**Acknowledgments** I would like to thank Elana Shohamy and Iair G. Or for comments and suggestions during the preparation of this chapter.

## Cross-References

► Assessing Multilingual Competence
► Assessing Second/Additional Language of Diverse Populations

- ▶ Cognitive Aspects of Language Assessment
- ▶ Critical Language Testing
- ▶ Washback, Impact, and Consequences Revisited

## Related Articles in the Encyclopedia of Language and Education

Sheena Gardner, Aizan Yaacob: Role Play and Dialogue in Early Childhood Education. In Volume: Discourse and Education

Christine Hélot: Awareness Raising and Multilingualism in Primary Education. In Volume: Language Awareness and Multilingualism

Kathryn M Howard: Language Socialization and Language Shift Among School-Aged Children. In Volume: Language Socialization

Amy Kyratzis, Marjorie Goodwin: Language Socialization and Social Practices in Children's Peer Interactions. In Volume: Language Socialization

## References

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English language learners. *Applied Measurement in Education, 27*(4), 261–272.

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm.* Smarter Balanced Assessment Consortium. Available at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Accomodations-for-under-represented-students.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological measurement*. Washington, DC: American Educational Research Association/American Psychological Association/National Council on Measurement in Education.

Bailey, A.L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly, 48*(3), 480–506.

Bailey, A. L., & Carroll, P. (2015). Assessment of English language learners in the era of new academic content standards. *Review of Research in Education, 39*, 253–294.

Bailey, A. L., & Heritage, M. (2008). *Formative assessment for literacy, Grades K-6: Building reading and academic language skills across the curriculum*. Thousand Oaks: Corwin/Sage Press.

Bailey, A. L., & Osipova, A. V. (2016). *Children's multilingual development and education: Fostering linguistic resources in home and school contexts*. Cambridge, UK: Cambridge University Press.

Bailey, A. L., Heritage, M., & Butler, F. A. (2014). Developmental considerations and curricular contexts in the assessment of young language learners. In A. J. Kunnan (Ed.), *The companion to language assessment*. Hoboken: Wiley-Blackwell.

Bailey, A. L., Blackstock-Bernstein, A., Ryan, E., & Pitsoulakis, D. (2016). Data mining with natural language processing and corpus linguistics: Unlocking access to school-children's language in diverse contexts to improve instructional and assessment practices. In S. El Atia, O. Zaiane, & D. Ipperciel (Eds.), *Data mining and learning analytics in educational research*. Malden: Wiley-Blackwell.

Barrueco, S., Lopez, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish-English bilingual preschoolers: A guide to best approaches and measures*. Baltimore: Brookes Publishing Company.

Beck. (1994). *Development of the Alchini Bizaad comprehension test of Navajo and English for young children.* Unpublished manuscript, University of Northern Arizona.
Bialystok, E., McBride-Chung, C., & Luk, G. (2005). Bilingualism, language proficiency and learning to read in two writing systems. *Journal of Educational Psychology, 97*(4), 580–590.
Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5–31.
Butler, F. A., & Stevens, R. (1997). Oral language assessment in the classroom. *Theory Into Practice, 36*(4), 214–219.
Cazden, C. (2001). *Classroom discourse: The language of teaching and learning* (2nd ed.). Portsmouth: Heinemann.
Chamot, A. U. (2005). The cognitive academic language learning approach (CALLA): An update. In P. Richard-Amato & M. A. Snow (Eds.), *Academic success for English language learners*. White Plains: Longman.
Cocking, R. R., & McHale, S. (1981). A comparative study of the use of pictures and objects in assessing children's receptive and productive language. *Journal of Child Language, 8*, 1–13.
Conti-Ramsden, G., & Durkin, K. (2012). Language development and assessment in the preschool period. *Neuropsychology Review, 22*(4), 384–401.
Council of Chief State School Officers. (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: CCSSO.
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.
Council of Europe. (2011). *European Language Portfolio.* Available at http://www.coe.int/t/dg4/education/elp/Default_en.asp
Davidson, F. (1994). Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms'. *Language Testing, 11*(1), 83–95.
Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale Press.
De Houwer, A. (1998). By way of introduction: Methods in studies of bilingual first language acquisition. *International Journal of Bilingualism, 2*(3), 249–263.
Department for Education. (2016). *The national curriculum*. Available at https://www.gov.uk/national-curriculum
Dockrell, J. E., & Marshall, C. R. (2015). Measurement issues: Assessing language skills in young children. *Child and Adolescent Mental Health, 20*(2), 116–125.
Echevarria, J., Vogt, M., & Short, D. (2004). *Making content comprehensible for English Language Learners: The SIOP model*. Needham Heights: Allyn & Bacon.
Faltis, C. (2015). Language advocacy in teacher education and schooling. In M. Bigelow & J. Ennser-Kananen (Eds.), *The handbook of educational linguistics* (pp. 65–77). New York: Routledge.
Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the U.S. *Language Assessment Quarterly, 11*(4), 432–457.
Frey, N., & Fisher, D. (2003). Linking assessment with instruction in a multilingual elementary school. In C. A. Coombe & N. J. Hubley (Eds.), *Assessment practices*. Alexandria: TESOL, Inc.
García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Oxford, UK: Wiley-Blackwell.
Gottardo, A., Chiappe, P., Yan, B., Siegel, L., & Gu, Y. (2006). Relationships between first and second language phonological processing skills and reading in Chinese-English speakers living in English-speaking contexts. *Educational Psychology, 26*(3), 367–393.
Graddol, D. (2006). *English next: Why global English may mean the end of English as a foreign language*. London: The British Council.
Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 22*, 337–354.

Heritage, M. (2013). Gathering evidence of student understanding. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 179–196). Thousand Oaks: Sage.
Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49*(1), 4–14.
Inbar, O., Shohamy, E., & Gordon, C. (2005). Considerations involved in the language assessment of young learners. *ILTA Online Newsletter, 2*, 3.
Jang, E. E. (2014). *Focus on assessment*. Oxford, UK: Oxford University Press.
Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational Researcher, 42*(4), 223–233.
Mahon, M., Crutchley, A., & Quinn, T. (2003). New directions in the assessment of bilingual children. *Child Language Teaching and Therapy, 19*, 237–243.
Masters, G., & Forester, M. (1996). *Developmental assessment*. Melbourne: Australian Council for educational Research Ltd.
McKay, P. (2005). Research into the assessment of school-age language learners. *Annual Review of Applied Linguistics, 25*, 243–263.
McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press.
Michel, M. C., & Kuiken, F. (2014). Language at preschool in Europe: Early years professionals in the spotlight. *European Journal of Applied Linguistics, 2*(1), 1–26.
National Association of the Education of Young Children (NAEYC). (2009). *Where we stand on assessing Young English-Language Learners*. Washington, DC: National Association of the Education of Young Children (NAEYC). Available at http://www.naeyc.org/files/naeyc/file/positions/WWSEnglishLanguageLearnersWeb.pdf.
National Educational Goals Panel (NEGP). (1998). In L. Shepard, S. L. Kagan, & E. Wurtz (Eds.), *Principles and recommendations for early childhood assessments*. Washington, DC: National Educational Goals Panel (NEGP).
National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.
National Research Council. (2001). In B. Bowman, M. S. Donovan, & M. S. Burns (Eds.), *Eager to learn: Educating our preschoolers*. Washington, DC: Committee on Early Childhood Pedagogy/National Academies Press.
NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
Nikolov, M. (2016). *Assessing young learners of English: Global and local perspectives*. Berlin: Springer.
Office of Head Start. (2010). *The head start child development and learning framework: Promoting positive outcomes in early childhood programs serving children 3–5 years old*. Arlington: Administration for Children and Families, U.S. of Health and Human Services.
Pellerin, M. (2012). Digital documentation: Using digital technologies to promote language assessment for the 21st century. *OLBI Working Papers*, *4*, 19–36.
Peña, E. D., & Quinn, R. (1997). Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools, 28*(4), 323–332.
Peña, E. D., Iglesia, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech- Language Pathology, 10*, 138–154.
Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and Non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10–28.
Pitsoulakis, D., & Bailey, A. L. (2016). *Extending learning progressions to self-assessment: Students finding "best fit" for next-steps learning and instruction.* Paper presented at the annual

meeting of the American Educational Research Association, Annual Conference, Washington, DC. Available at www.dllp.org

Powell, D. R., & Sigel, I. E. (1991). Searches for validity in evaluating young children and early childhood programs. In B. Spodek & O. N. Saracho (Eds.), *Issues in early childhood education* (Yearbook in Early Childhood Education, Vol. 2, pp. 190–212). New York: Teachers College Press.

Puckett, M. B., & Black, J. K. (2000). *Authentic assessment of the young child*. Upper Saddle River: Prentice Hall.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing, 18*, 429–462.

Rea-Dickins, P., & Rixon, S. (1997). The assessment of young learners of English as a foreign language. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language and education* (Language Testing and Assessment, Vol. 7, pp. 151–161). Dordrecht: Kluwer.

Scollon, R., & Scollon, S. (1981). *Narrative, literacy, and face in interethnic communication*. Norwood: Ablex.

Silburn, S. R., Nutton, G., McKenzie, J., & Landrigan, M. (2011). *Early years English language acquisition and instructional approaches for Aboriginal students with home languages other than English: A systematic review of the Australian and international literature*. Darwin: The Centre for Child Development and Education, Menzies School of Health Research.

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education, 39*(1), 215–252.

Tabors, P. (2008). *One child, two languages: A guide to preschool educators of children learning English as a second language* (2nd ed.). Baltimore: Brookes Publishing.

Téllez, K., & Mosqueda, E. (2015). Developing teachers' knowledge and skills at the intersection of English language learners and language assessment. *Review of Research in Education, 39*(1), 87–121.

Tsagari, C. (2004). Alternative assessment: Some considerations. *Language Testing Update, 36*, 116–125.

Uccelli, P., Barr, C. D., Dobbs, C. L., Galloway, E. P., Meneses, A., & Sanchez, E. (2014). Core academic language skills: An expanded operational construct and a novel instrument to chart school-relevant language proficiency in preadolescent and adolescent learners. *Applied Psycholinguistics, 36*, 1–33.

Vellutino, F. R., & Scanlon, D. M. (2001). Emergent literacy skills, early instruction, and individual differences as determinants of difficulties in learning to read: The case for early intervention. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 295–321). New York: Guilford Press.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach* (pp. 143–161). Basingstoke: Palgrave Macmillan.

Wiliam, D. (2006). Commentary: Formative assessment: Getting the focus right. *Educational Assessment, 11*(3–4), 283–289.

# Assessing Second/Additional Language of Diverse Populations

Constant Leung and Jo Lewkowicz

**Abstract**

In this chapter we address second/additional language assessment from two perspectives. First, we look at assessment of linguistic minority students within a national context where their second/additional language is the predominant majority language or an important auxiliary language in the society. Then we look at other contexts where bi/multilingualism is strongly encouraged (e.g., in countries within the European Union) and second/additional language assessment plays an important role in determining educational success. In both cases, for reasons of our professional experience, we focus our discussion on English as the second/additional language, though we refer to other languages where appropriate. We present in turn the major developments related to each of the themes and consider some of the problems and difficulties associated with developing assessment frameworks that are appropriate for such diverse populations and contexts. We detail some of the advances in Europe as we believe they signal some of the likely future directions reflecting progressive societal recognition of the value of a person's proficiencies in different languages within their linguistic repertoires. We also highlight the dangers of using language assessment criteria modeled on one kind of population for another inappropriately for reasons of administrative and policy expediency and context-insensitive public accountability.

C. Leung (✉)
Centre for Language, Discourse and Communication, School of Education, Communication and Society, King's College London, London, UK
e-mail: constant.leung@kcl.ac.uk

J. Lewkowicz
University Council for the Certification of Language Proficiency, University of Warsaw, Warszawa, Poland
e-mail: jlewkowicz@uw.edu.pl

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_23

**Keywords**
Additional language • Linguistic minority • Linguistic diversity

## Contents



## Introduction

With increasing movements of people across international boundaries and the unabated spread of plurilingualism into national education systems, intergovernmental cooperation, and multinational business enterprises, the teaching and assessment of second/additional language proficiency have continued to be a major item on the educational agenda in many world locations. In this chapter we focus on (1) additional/second language assessment of linguistic minority students in a context where this language is the predominant majority language or an important auxiliary language in society and (2) language assessment designed to track language development of learners of an additional/second/foreign language in other diverse contexts. These two themes will be discussed in two parallel sections on developments and problems identified to date. Although we will be mainly dealing with English as an additional, second, or foreign language (EAL/ESL/EFL), our discussion will refer to work in other languages where appropriate. We will also be referring to some recent developments in Europe which, in our view, signal some of the likely future directions reflecting progressive societal recognition of the value of a person's proficiencies in different languages. "Assessment" is used as a superordinate term throughout this discussion to refer to all forms of assessment, including standardized tests. In the case of English language, the terms "additional language (AL)" and "second language (SL)" broadly share the same meaning in contexts where English is learned and used by learners from diverse language (non-English speaking) backgrounds; in the United States, the term "English language (learners)" is also used. In this discussion we will use "additional language" as a generic term

and explicitly signal foreign language and modern language where specificity of meaning would warrant their use.

## Theme 1: Assessing AL as a Distinctive Curriculum Phenomenon

### Early Developments

The need for supporting EAL students is not new; it has, however, become increasingly important with the growth of mobility throughout the world and the steady rise in EAL student populations in English-speaking countries. For instance, the percentage of English language learners in the US public school population grew from 8.7% (4.1 million students approx.) in 2002/2003 to 9.1% (4.4 million) in 2011/2012 (National Center for Education Statistics 2015). Similarly, the numbers have been increasing in England – in 2005, EAL students constituted 11.7% of the total population in elementary schools and 9.1% in secondary schools (DfES 2005a) rising to 18.7% (612,000 approx.) and 14.3% (436,000 approx.), respectively, in 2014 (DfE 2014). Since the early 1990s, there have been two related but, paradoxically, opposite developments in the assessment of the additional language development of linguistic minority students in English-speaking countries. In a number of education jurisdictions, there has been a major effort to develop distinctive EAL assessment frameworks; at the same time, many national systems have adopted an inclusive policy and practice of putting all students through large-scale standardized public assessment schemes without distinction. These two opposing developments will be discussed in turn.

There has been a growing awareness on the part of some educators and policy makers that additional language development in the context of mainstream schooling and social participation is different from first language development and foreign language learning (e.g., learning French as a subject in an English-medium school curriculum). EAL students enter their local school system at different ages and with varying background in English language learning. Learning English can add considerable demand to the academic challenges faced by individual students. (The same can be said for any educationally or societally dominant auxiliary language in any part of the world; see Shohamy 2007, for a wider discussion.) For this reason a good deal of effort has gone into the systematic development of dedicated EAL assessment frameworks across a number of education jurisdictions.

### Major Contributions

One of the first of such attempts is the National Language and Literacy Institute of Australia (NLLIA) framework (McKay 1992, 2007) which sets out to provide grade-level classroom-based EAL assessment descriptors. The descriptors take into account the use of English for subject content learning in ordinary classroom

contexts. Other Australian assessment frameworks include the curriculum-oriented ESL Scope and Scales in South Australia (SACSA undated) and the EAL/D Learning Progression: Foundation to Year 10 (ACARA 2014) which set out the stages of progression related to the Australian national educational standards. The professional association Teachers of English to Speakers of Other Languages (TESOL 1997, 2006) in the United States has produced the K-12 English Language Standards which have been designed to provide teachers with broad requirements of EAL development for social and academic purposes at different stages of schooling. Teachers are encouraged to use these descriptors to generate contextualized local EAL assessment criteria.

We have also seen that the policy move toward greater central control, public accountability, and economic rationalism, which originated in the 1970s, has been further consolidated in many countries. Over 15 years ago, Broadfoot and Pollard (2000, p.13) captured this powerful trend thus:

> The underlying rationale here [emphasizes] the beneficial role of market forces and competition in driving up standards, and controlling 'producer interests' … In such a model, assessment and measurement has a particular role in providing 'objective' information on which educational 'consumers' such as parents and governments can base their decisions.

The tendency to regard school and university education as producers of skilled and knowledgeable labor is now commonplace in many parts of the world. However, the powerful central control of the accountability mechanisms (e.g., mandatory school inspections), once exclusively the domain of national governments, is increasingly being augmented by supranational quasi-governmental bodies such as the Organization for Economic Cooperation and Development. The evaluation programs run by these organizations are presented as politically neutral, and they yield enormous influence on national policy discourse and formation. The "standards" and evaluation methodology adopted by these programs are largely driven by economic demands and labor market considerations; an obvious example is the Program for International Student Assessment (PISA). PISA data and evaluations are adopted by national governments to justify their policy decisions. Educational accountability is beginning to be shaped by "an emerging regime of global educational governance" (Meyer and Benavot 2013, p. 11).

Many education systems have adopted the use of standard-based assessment and public reporting of student performance as part of policy implementation and monitoring. The current legislation connected to the No Child Left Behind (NCLB) policy in the United States, for example, requires regular assessment and reporting of results for all school students. Similar statutory requirements exist in places such as Australia and England. Additional language education has not been exempted from this process. Proponents of this approach argue that this "common treatment" contributes to social integration and educational equal opportunities (e.g., Travers and Higgs 2004; cf Menken 2008). The problems with adopting a standard-based assessment approach that does not take account of linguistic diversity will be addressed next.

## Problems and Difficulties

The use of assessment to promote a particular kind of public policy is not, however, unproblematic in terms of potential misuse of assessment. Such issues are readily apparent in the high-stakes, standardized tests of language and content subjects adopted in the United States. These have been developed for monolingual English speakers but are used to assess all students, including English language learners (Menken, Hudson, and Leung 2014). Although some accommodations may be available to the emergent bilinguals, these vary from state to state and are not uniformly available. For example, translation of tasks may only be provided for the more common home/first languages spoken by these students. Furthermore, the underlying assumption of providing such accommodation is that the emergent bilinguals have appropriate schooling in their home languages. However, if the student has not learned the content covered by the test in English, then translation of the task is unlikely to help. As Menken et al. (2014, p. 606) point out, "the language of instructions must match the language of the test for scores to be valid."

English language learners in the United States are in addition required to demonstrate improvement in some form of standardized English proficiency examination. However, the test used varies from state to state, and given that each operationalizes the academic language it is meant to assess variably, it is not surprising that the validity and reliability of some of these tests have been questioned by researchers (e.g., Bailey and Huang 2011). Although the tests are meant to predict English language learners' readiness to participate together with their monolingual peers, there is evidence to suggest that performing well on an English proficiency test does not necessarily lead to success on the English arts (content) examination (and vice versa).

Similar issues can also be seen in other places. The following is an illustrative example from recent English experience. The government-sponsored assessment of English (the term "literacy" is used sometimes) is based on the system-wide rating scales to be applied to "everyone." The statutory assessment of elementary and lower secondary students in England requires the National Curriculum assessment criteria to be applied to all, irrespective of first or second language backgrounds; schools were advised that:

> Summative assessment for bilingual [EAL] pupils, as for all pupils, should be based on national curriculum measures . . .. It is not recommended that additional locally developed scales of fluency are used . . .' (DfES 2005b, p. 6)

The use of common assessment criteria for all students without exception may be justifiable on grounds of an "inclusive" approach to education. Here "inclusiveness" is taken to mean common educational treatment irrespective of differences in language backgrounds (see Leung 2001, 2009 for a detailed discussion). In terms of usefulness of assessment outcome, however, the appropriateness of using first language development models for the assessment of additional language development is questionable. For instance, in the English (subject) National Curriculum, the

attainment target for Level 4 Speaking and Listening (expected level of attainment for 11/12-year-olds) was as follows:

> Pupils talk and listen with confidence in an increasing range of contexts. Their talk is adapted to the purpose: developing ideas thoughtfully, describing events and conveying their opinions clearly. In discussion, they listen carefully, making contributions and asking questions that are responsive to others' ideas and views. They use appropriately some of the features of standard English vocabulary and grammar. (DfES and QCA 1999, p. 55)

This attainment target statement provided, arguably, a reasonably workable general description of the range and kinds of spoken language use for school purposes by first language speakers. Note that not only was the student expected to use spoken English to engage in a range of academic activities, s/he was also expected to do it in socioculturally acceptable ways; qualifiers such as "with confidence," "thoughtfully," and "appropriately" all point to the sort of language repertoire expected of someone who has had substantial exposure and use of English in a native/first language speaking environment. This level description would not even begin to make sense for either summative or formative purposes in the case of a 12-year-old beginner learner of English, say, from Poland or Somalia. Yet, if we turn to a lower level (younger age) description, there would be an odd sense of misfit because of the age and maturation factors built into first language scales. The Level 1 description for Speaking (threshold, officially EAL friendly), for instance, was as follows:

> Pupils speak about matters of immediate interest in familiar settings. They convey meaning through talk and gesture and can extend what they say with support. Their speech is sometimes grammatically incomplete at word and sentence level. (QCA 2000, p. 13)

Here the level description was clearly modeled on a much younger child, about the age of 5 or 6, who might be happy to engage with others in an uninhibited manner. A 12-year-old English language beginner would be unlikely to talk about matters of immediate interest in a secondary school setting. The greatest challenge for such a student is likely to be finding the necessary vocabulary and phrases, grammatically complete or not. This example shows that assessment criteria which have been developed with first language development norms and assumptions can be conceptually ill fitting and, worse, misleading in terms of the assessment outcome yielded. This insistence on "first language for all" policy has persisted despite recent changes in other aspects of statutory assessment in England.

In the content areas, a similar situation exists. EAL students, irrespective of their English language proficiency and schooling backgrounds, are expected to participate in standardized subject assessment which has been devised with native speakers in mind. For those students who are still learning to use English for academic purposes effectively, the English language in standardized assessments can pose an additional linguistic challenge that distorts their ability to demonstrate their content knowledge. Whenever this happens it would make the test scores "invalid as indicators of content knowledge and achievement" (Butler and Stevens 2001, p. 411). In a study

comparing test performance of monolingual and emergent bilingual school students across the United States, Menken (2008) found a widening gap with age for both reading and math suggesting that a lack of English language proficiency can have a delirious effect on test performance. Menken et al. (2014, p. 605) further argue that "language is a liability for emergent bilinguals for whom testing is primarily punitive in outcome." All of this raises serious fundamental questions about the validity of using a set of non-differentiated criteria for the assessment of additional language students' English and curriculum achievements.

## Theme 2: Assessing Additional/Foreign Language Development of Learners in Diverse Contexts

### Early Developments

In contrast to theme 1 which looks at issues related to multilingualism within national contexts, this theme focuses on issues related to assessing diverse languages ensuring comparability of measures. Within Europe it has become increasingly important to identify and recognize "the kinds of language proficiency needed by European citizens to interact and cooperate effectively" (Figueras et al. 2005, p. 263). This has been facilitated by a number of Council of Europe initiatives, starting in 1957 with the first intergovernmental conference on European cooperation in language teaching. A significant early development was the Council of Europe's publication in 1975 of the Threshold Level, "the specification in operational terms of what a learner should be able to do when using the language interactively." The 1990s subsequently saw the specification of intermediate- (Waystage) and higher-level (Vantage) objectives and the development of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001). The CEFR now provides a yardstick against which language proficiency in the different languages of Europe can be described.

### Major Contributions

The CEFR has undoubtedly had a major impact not only on the teaching and learning of languages but also on languages assessment, facilitating "comparisons between different systems of quantifications" (Council of Europe 2001, p. 21). It was initially envisaged as a pan-European framework for a linguistically diverse continent within which there is free movement of peoples. Its influence, however, has far exceeded initial expectations, and the CEFR is now a supranational document translated into 40 languages (North 2014), impacting on educational language policy worldwide (see Byram and Parmenter 2012 for a discussion of its use in a range of contexts). The "A1" (lowest) to "C2" (highest) CEFR levels are routinely referenced by language course and test providers as well as textbooks. (For details of the CEFR, see Council of Europe 2001.)

There have been numerous initiatives to address some of the underlying issues with the CEFR. A primary concern has been the abstract nature of the document and the difficulty of applying the framework, particularly at adjacent levels. To this end the Council of Europe has developed a Manual outlining the stages in the process of relating specific language tests to the CEFR (Council of Europe 2005; 2009a) as well as a Reference Supplement (Council of Europe 2009b) that provides technical support for this purpose. These documents have facilitated one of the primary objectives of the CERF to be attained – comparability across different languages and levels of modern languages assessment – and there are numerous reports in the literature presenting alignment projects across a range of languages and contexts within and beyond Europe (see, among others, Bechger et al. 2009, for Dutch as a second language; Wu and Wu 2010, for an English test in Taiwan).

A related and very significant development in the assessment of languages in Europe was the introduction of the European Language Portfolio (ELP). The portfolio was inspired by the Council of Europe as a result of the 1991 Rüschlikon Symposium (see Little 2002 for details). It was introduced at a time when not only was the number of users of English in Continental Europe growing, but when the need for recognizing multilingualism in the United Kingdom was rising (King 2001).

The ELP "is a personal document designed to make the language learning process more transparent to the learner and to report an individual's achievement at any level and in any language in an internationally transparent way." It is one of five documents that makes up the Europass, a way of making a person's "skills and qualifications clearly and easily understood in Europe" for the benefits of citizens, employers, and education providers (https://europass.cedefop.europa.eu/en/about).

A number of key features distinguish ELP from other means of language assessment. Importantly, it promotes equal recognition of all languages learned by individuals, utilizing a model developed after extensive piloting across 15 countries within the Council of Europe (Little 2005). The established model is one that has a standardized language passport for all adults but that allows for variability within the different sections of the portfolio, thus reflecting the Council of Europe's ideal of "unity in diversity" (Little 2002, p. 184).

ELP has gained widespread acceptance in Europe. Initially it was developed for use with adults, but it has been successfully adapted for adolescents and young learners, taking account of their specific needs and interests (see Hasselgreen 2005, for an account of one such project). In 2011, after 118 ELPs had been validated, the system of validation was modified to one of "online registration based on the principle of self-declaration" (Little et al. 2011, p. 16).

The formative nature of ELP assessment allows individuals to engage in the portfolio process from an early age and to continue updating its contents as its owner perceives necessary. Designed to supplement official certificates awarded through formal education, it allows its owner to demonstrate any language learning that has taken place outside the formal educational setting, e.g., within a bilingual home or while traveling abroad.

Comparability across languages within a single portfolio or across different portfolios is made possible through the CEFR. The CEFR also provides a means for self-assessment which is an integral part of the portfolio. This central aspect of the ELP is believed to promote self-reflection. In addition, "learners gain 'insider' access to the processes of 'social moderation' that underlie the CEFR's common reference levels and to the interaction between curriculum and assessment that is fundamental to any worthwhile educational enterprise" (Little 2005, p. 335).

A further European development which recognizes the need for assessment across the diverse languages of Europe is that of DIALANG (http://www.lancaster.ac.uk/researchenterprise/dialang/). This is a low-stakes, computer-based, and internet-delivered test of reading, writing, listening, grammatical structures, and vocabulary covering all six levels articulated in the CEFR. The test can be used by individuals wanting to assess their language level in one of 14 languages or by institutions for diagnostic or placement purposes. It is readily and freely available to users, providing them with immediate feedback on their performance and with information on how they can improve their proficiency. (For details see Alderson and Huta 2005.)

## Problems and Difficulties

Despite ongoing attempts to make the CEFR more accessible to end users, there continue to be a number of inherent flaws in the system that need to be addressed if the ideal of comparability and transparency across languages and levels of proficiency is to be fully accomplished. Although levels are specified in "Can do" terms, there appears to be a considerable lack of clarity as to how these should be interpreted and operationalized by end users (Goodier 2014). Moreover, there is no indication as to how adequate each performance needs to be to qualify for a particular level. It is often difficult to distinguish between two adjacent proficiency levels such as, for example, between "Can understand a wide range of demanding longer texts" (C1) and "Can understand with ease virtually everything heard or read" (C2). How long are longer texts and how wide ranging do they need to be to qualify for C1 rather than for C2?

Another shortcoming of the CEFR is that its specifications do not detail the nature of tasks that would be appropriate for each level or account for the development of cognitive and metacognitive processing as one progresses from one level to the next. (For a comprehensive discussion of these and related issues, see Weir 2005.) The reporting of a single, global level such as B1 or B2 would appear somewhat simplistic, viewing language as unidimensional and ignoring the complexity of language development across the different skills (see Harsch 2014, for a fuller discussion). Furthermore, it is becoming quite clear that the language competence model underlying the CEFR needs updating. Recent research in the use of English in linguistically diverse contexts has shown that the linguistic and sociolinguistic norms embedded in the framework and its rating scales need revising and extending

to account for contemporary conditions (see Jenkins and Leung 2013, this volume; Leung 2014; Leung and Lewkowicz 2012).

## Work in Progress

Demographic, political, and social developments in the past 50 years suggest that linguistic diversity continues to spread and intensify in many societies across the world. Given the unquestionable influence of the CEFR on assessment within Europe and beyond, it is not surprising that the framework has been the focus as well as the basis of much ongoing research. The publication of the Manual (Council of Europe 2009a) has facilitated ongoing work on test alignment and on demonstrating ways in which a given test meets a specific level, yet it is probable that some of the more local alignment projects are not being reported as extensively as those conducted by the major test providers. Nevertheless, all attempts at alignment are contributing to our understanding of what learners can do at the various levels of the framework, thus helping to refine our understanding of language proficiency (Harsch 2014). At the same time, alignment projects are informing and being informed by the growing number of learner corpora, such as the Cambridge Learner Corpus which is based on samples of learner English from student exam scripts (http://www.cambridge.org/elt/corpus). Attempts are also underway to "suggest improvements for the CEFR based on research outcomes." These are the objectives of the Second Language Acquisition and Testing in Europe (SLATE) network, a group of European-wide researchers undertaking a range of projects (for details see (www.slate.eu.org)) to better understand how language develops over the six levels of the CEFR. Enhancing understanding of CEFR levels through empirical study is also the aim of the English Profile Program which is working toward a profile of the grammar, vocabulary, and linguistic functions characteristic at each CEFR level (see www.englishprofile.org).

In more specific national contexts, the design and implementation of language assessment for educational, professional, and other purposes needs to pay attention to diverse language backgrounds of speakers and contexts of use. The need to take account of diversity in public education has been recognized by many educational jurisdictions. However, the translation of this recognition into action is still at an early stage. In the United States, for instance, there have been various attempts to provide accommodations for English language learners in formal examinations and tests by allocating additional time or using bilingual material. Research on the impact of such accommodations has so far been inconclusive (see Abedi et al. 2004; Menken 2008; among others). More recently under pressure of the mandatory yearly assessment stipulated by the No Child Left Behind legislation, the assessment framework for English language proficiency of school students from linguistic minority backgrounds developed by a large multistate consortium, World-Class Instructional Design and Assessment (WIDA, https://www.wida.us/assessment/), takes account of language use and language demands in content areas such as

Math and Science. The increasing adoption of the Common Core State Standards (which set out school subject content specifications) across the United States has amplified the need for content assessment in all subjects to consider urgently the by now well-understood disadvantages of monolingually conceived assessment instruments for students of diverse language backgrounds (Menken, Hudson and Leung 2014; Solórzano 2008). In many ways the situation in the United States resonates with developments in other places where assessment frameworks and practices are beginning to embrace the needs of linguistically diverse students. For instance, in England where the school system has long recognized linguistic diversity, the statutory assessment system has continued to be monolingually oriented. A content-sensitive English as an additional language assessment framework for schools is just beginning to be developed through an independent project (Evans et al. 2015). Similarly, an initiative involving state-level education units and academic institutions to develop diversity-aware pedagogy and diagnostic assessment has recently been established in Germany (see http://www.biss-sprachbildung.de/).

The use of assessment to promote learning has been receiving increasing attention in recent years in all areas of education (e.g., Assessment Reform Group 2002; Black and Wiliam 1998, 2009; Swaffield 2008; Wiliam 2011, among others). In the field of additional language assessment, there is now an established body of work that relates assessment to pedagogy. Given that the primary focus of this approach to assessment is to facilitate learning, a good deal of the research and development is focused on curriculum and classroom practices. The term "formative assessment" (otherwise known as Assessment for Learning) is often used to cover work in this energetic field (e.g., see the collection in Davison and Leung 2009). Some educational jurisdictions have been developing their hitherto summatively oriented assessment frameworks to include an element of formative assessment. For instance, the Hong Kong School-Based Assessment component of the public school-leaving examinations (that includes English language as a subject) states that:

> SBA emphasises the assessment of a wide range of abilities which offers a comprehensive appraisal of students' performance. By integrating learning and teaching with assessment, it helps students understand their strengths and weaknesses through quality feedback from teachers. SBA also reduces dependence on the results of public examinations and boosts students' confidence and motivation to learn and enhances autonomous learning. (http://www.hkeaa.edu.hk/en/sba/introduction/)

Another example is the promotion of formative assessment within the official curriculum framework in Wales where Welsh as an additional language is part of the national curriculum (Llywodraeth Cynulliad Cymru 2010). The National Certificates of Educational Achievement assessment framework in New Zealand has a portfolio component that comprises students' own collection of evidence of learning (http://seniorsecondary.tki.org.nz/Learning-languages/What-s-new-or-different/). For language subjects such as French or Japanese, students are asked to collect visual/audio samples of unrehearsed spoken and written language use in and out of school contexts. This approach aims to generate student-led enquiry and assessment with

clear formative benefits (Absolum et al. undated). This kind of system-wide attempt to use assessment to promote learning has been accompanied by a growing body of research that seeks to expand and refine the concepts and theories involved (see Lantolf and Poehner 2013; Turner and Purpura 2016; among others). International interest in such initiatives is very high with developments being followed closely by practitioners and researchers alike.

## Future Directions

Research and professional development now needs to address issues related to changes in demographics and in language practices. As linguistic diversity accelerates the fundamental question of what constitutes language competence (in English or any other language used as the medium of instruction) in relation to curriculum participation is fast becoming a central issue in terms of access to mainstream educational provision and effective learning. Where students are required to participate in the full range of curriculum subjects and mandatory assessment, it is necessary to ask the following language model-cum-construct-related questions: How should assessment deal with the relationship between curriculum content and classroom language use? What is language proficiency in curriculum and schooling contexts? While some education jurisdictions have taken the first steps in addressing these issues (as indicated above), others have barely begun (see, e.g., European Commission 2013).

Another issue that needs to be addressed particularly within the context of young learners is how knowledge of all languages can be simultaneously valued. As Shohamy (2014, p. 16) points out in relation to the complex languages situation in Israel, schools too often try to "turn multilingual realities into monolingual islands" for political reasons. This tendency is also evident in the European Union. Given the virtually unrestricted freedom of movement within the EU and the high levels of immigration into Europe, many classrooms at all levels of education have students from a range of different L1 backgrounds and with variable knowledge of additional languages. This inevitably poses questions about teachers being able to assess and cater for the language needs of all their linguistically diverse students in terms of promotion of additional language learning and bi/multilingualism. Teachers require training in assessment, particularly formative assessment, which is all too often lacking (Leung 2013; Rea-Dickins 2000, 2001).

The ELP and other similar packages will have to take account of the growing use of English as lingua franca (ELF), particularly in the higher education context where an increasing number of courses are being offered in English (Coleman 2006; Jenkins and Leung 2014). ELF is a commonplace phenomenon (see Jenkins 2006, 2014; Seidlhofer 2011) often evident in contexts where the majority, if not all, of the participants are nonnative speakers of English and where Anglophone sociolinguistic rules do not necessarily apply. Thus, any language assessment designed for transnational use needs to pay attention to the resultant emergent forms and practices. The challenge in the near future will be to address these issues so that the

CEFR can be applied in the way aspired to by the Council of Europe, achieving transparency, consistency, and uniformity across languages and levels of proficiency.

## Concluding Remarks

This discussion has focused on key conceptual and design issues related to additional language learners and users in formal education and other settings. Although there are contextual differences, there is a broad common goal – to build systems and frameworks that can effectively represent language learners' achievement in a world of increasing population mobility. Assessment frameworks such as the CEFR and WIDA are examples par excellence. At the same time, in the public education domain, the international policy trend toward public accountability has introduced many conceptual challenges. From the point of view of this discussion, a key issue is the use language assessment criteria modeled on one kind of population being used for another. In many mainstream education contexts, the problems largely arise from using first language descriptors for assessing additional language performance, while in the European context, adult-oriented language descriptors are being used as a model for assessing young learners. Furthermore, as language practices change, there is a need for assessment criteria, descriptors, and rating scales to be revised periodically to reflect emergent forms of use and competence. Progress toward resolving this kind of fundamental issue is likely to require both technical and conceptual development through systematic research and some form of public policy realignment to accommodate diversity in assessment criteria.

## Cross-References


- ▶ Assessing English as a Lingua Franca
- ▶ Assessing English Language Proficiency in the United States
- ▶ Assessing Multilingual Competence
- ▶ Language Assessment in Indigenous Contexts in Australia and Canada
- ▶ The Common European Framework of Reference (CEFR)
- ▶ Utilizing Accommodations in Assessment


## Related Articles in the Encyclopedia of Language and Education


Marcia Farr: Literacies and Ethnolinguistic Diversity: Chicago. In Volume: Literacies and Language Education

Ingrid Gogolin, Joana Duarte: Superdiversity, Multilingualism and Awareness. In Volume: Language Awareness and Multilingualism

Barbara Seidlhofer: English as Lingua Franca and Multilingualism. In Volume: Language Awareness and Multilingualism

Inge Sichra: Language Diversity and Indigenous Literacy in the Andes. In Volume: Literacies and Language Education

Massimiliano Spotti, Sjaak Kroon: Multilingual Classrooms at Times of Superdiversity. In Volume: Discourse and Education

## References


Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Absolum, M., Flockton, L., Hattie, J., Hipkins, R., & Reid, I. (Undated). Directions for assessment in New Zealand. Retrieved 7 Oct 2015, from http://www.tki.org.nz/r/assessment/research/mainpage/directions/

Alderson, J. C., & Huta, A. (2005). The development of a suite of computer based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320.

Assessment Reform Group. (2002). Assessment for learning: 10 principles. Retrieved 1 Apr 2014, from http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf

Australian Curriculum Assessment and Reporting Authority. (2014). *English as an additional language or dialect: Teacher resource – EAL/D learning progression: Foundation to Year 10*. Sydney: ACARA.

Bailey, A., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing, 28*(3), 343–365.

Bechger, T., Kuijper, H., & Maris, G. (2009). Standard setting in relation to the Common European Framework for Languages. The case of the state examination of Dutch as a second language. *Language Assessment Quarterly, 6*(2), 126–150.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–73.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31.

Broadfoot, P., & Pollard, A. (2000). The changing discourse of assessment policy: The case of English primary education. In A. Filer (Ed.), *Assessment: Social practice and social product* (pp. 11–26). London: RoutledgeFalmer.

Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas. *Language Testing, 18*(4), 409–427.

Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference. The globalization of language education policy*. Bristol: Multilingual Matters.

Coleman, J. (2006). English-medium teaching in European higher education. *Language Teaching, 39*(1), 1–14.

Council of Europe. (2001). *Common European framework for reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2005). *Relating language examinations to the common european framework of reference for languages: Learning, teaching, assessment (cef). Manual: Preliminary pilot version. Dgiv/edu/lang 2003, 5*. Strasbourg: (Language Policy Division) Council of Europe.

Council of Europe. (2009a). *Relating language examinations to the Common European Framework for Languages (CEF). A manual*. Strasbourg: Language Policy Division.

Council of Europe. (2009b). *Reference supplement to the Manual for relating language examinations to the CEFR*. Strasbourg: Council of Europe: Language Policy Division.

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393–415.

Department for Education. (2014). Schools, pupils and their characteristics: January 2014. Retrieved 13 July 2015, from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/410543/2014_SPC_SFR_Text_v102.pdf

Department of Education Employment, & Qualifications and Curriculum Authority. (1999). *English – The national curriculum for England*. London: DfEE and QCA.

Department for Education and Skills. (2005a). Pupil characteristics and class sizes in maintained schools in England, January 2005 (provisional). Retrieved 25 July 2006, from http://www.dfes.gov.uk/rsgateway/DB/SFR/s000574/sfr16-2005.pdf

Department for Education and Skills. (2005b). *Aiming high: guidance on assessment of pupils learning English as an additional language*. Nottingham: DfES.

European Commission. (2013). *Study on educational support for newly arrived migrant children*. Brussels: Publications Office, European Union.

Evans, M., Jones, N., Leung, C., & Liu, Y. C. (2015). EAL assessment and evaluation framework. *NALDIC Quarterly, 15*(2), 4–7.

Figueras, N., North, B., Takala, S., & Verhelst, N. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*(3), 261–279.

Goodier, T. (2014). *Working with CEFR can-do statements An investigation of UK English language teacher beliefs and published materials.* Unpublished M.A. dissertation, King's College, London.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11*(2), 152–169.

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 22*(3), 337–354.

Jenkins, J. (2006). Current perspectives on teaching World Englishes and English as a Lingua Franca. *TESOL Quarterly, 40*(1), 157–181.

Jenkins, J. (2014). *English as a lingua franca: The international university*. London: Routledge.

Jenkins, J., & Leung, C. (2013). English as a Lingua Franca. In A. Kunnan (Ed.), *Companion to language assessment*. Hoboken: Wiley-Blackwell.

King, L. (2001). The European Year of Languages – Taking forward the languages debate. *Language Teaching, 34*(1), 21–29.

Lantolf, J. P., & Poehner, M. E. (2013). The unfairness of equal treatment: Objectivity in L2 testing and dynamic assessment. *Educational Research and Evaluation, 19*(2–3), 141–157.

Leung, C. (2001). English as an additional language: Distinctive language focus or diffused curriculum concerns? *Language and Education, 15*(1), 33–55.

Leung, C. (2009). Mainstreaming: Language policies and pedagogies. In I. Gogolin & U. Neumann (Eds.), *Streitfall Zweisprachigkeit – The bilingualism controversy* (pp. 215–231). Wiesbaden: VS Verlag für Sozialwissenschaften.

Leung, C. (2013). Classroom-based assessment: Issues for language teacher education. In A. Kunnan (Ed.), *Companion to language assessment* (Vol. III, pp. 1510–1519). Hoboken: Wiley-Blackwell.

Leung, C. (2014). Communication and participatory involvement in linguistically diverse classrooms. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and Bilingual Education* (pp. 123–146). New York: Routledge.

Leung, C., & Lewkowicz, J. (2012). Language communication and communicative competence: A view from contemporary classrooms. *Language and Education, 27*(5), 398–414.

Little, D. (2002). The European Language Portfolio: Structure, origins, implementation and challenges. *Language Teaching, 35*(3), 182–189.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing, 22*(3), 321–336.

Little, D., Goullier, F., & Hughes, G. (2011). *The European language portfolio: The story so far (1991–2011)*. Strasbourg: Council of Europe.

Llywodraeth Cynulliad Cymru. (2010). *How to develop thinking and assessment for learning in the classroom*. Cardiff: Welsh Assembly Government.
McKay, P. (1992). *ESL development: Language and literacy in schools project: Vol 1 and 2*. East Melbourne: National Languages and Literacy Institute of Australia (NLLIA).
McKay, P. (Ed.). (2007). *Assessing, monitoring and understanding English as a second language in schools: The NLLIA ESL Bandscales Verison 2*. Brisbane: Queensland University of Technology and Independent Schools Queensland.
Menken, K. (2008). *English learners left behind*. Clevedon: Multilingual Matters.
Menken, K., Hudson, T., & Leung, C. (2014). Symposium: Language assessment in standards-based education reform. *TESOL Quarterly, 48*(3), 586–614.
Meyer, H.-D., & Benavot, A. (2013). PISA and the globalization of education governance: Some puzzles and problems. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 7–26). Oxford: Symposium Books.
National Center for Education Statistics. (2015). English Language Learners. Retrieved 10 July 2015, from http://nces.ed.gov/programs/coe/indicator_cgf.asp
North, B. (2014). *The CEFR in practice*. Cambridge: Cambridge University Press.
Qualifications and Curriculum Authority. (2000). *A language in common: Assessing English as an additional language*. London: QCA.
Rea-Dickins, P. (2000). Current research and professional practice: Reports of work in progress into the assessment of young language learners. *Language Testing, 17*(2), 245–249.
Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing, 18*(4), 429–462.
Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford: Oxford University Press.
Shohamy, E. (2007). Reinterpreting globalization in multilingual contexts. *International Multilingual Research Journal, 1*(2), 1–7.
Shohamy, E. (2014). The weight of English in global perspective: The role of English in Israel. *Review of Research in Education, 38*(1), 273–289.
Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English Language Learners. *Review of Educational Research, 78*(2), 260–329.
South Australian Curriculum Standards and Accountability Framework (SACSA). (Undated). ESL Scope and Scales: SACSA, South Australia.
Swaffield, S. (Ed.). (2008). *Unlocking assessment: Understanding for reflection and application*. London: Routledge.
Teachers of English to Speakers of Other Languages (TESOL). (1997). *ESL standards for pre-k-12 students*. Alexandria: TESOL.
TESOL. (2006). *PreK-12 English Language Proficiency Standards*. Alexandria: Teachers of English to Speakers of Other Languages, Inc.
Travers, P., & Higgs, L. (2004). Beyond the naming of parts: Working with pupils at Key Stages 3 and 4 in the English curriculum. In P. Travers & G. Klein (Eds.), *Equal Measures: Ethnic minority and bilingual pupils in secondary schools* (pp. 27–44). Stoke on Trent: Trentham Books.
Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment*. Berlin: DeGruyter Mouton.
Weir, C. J. (2005). Limitations of the common european framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.
Wiliam, D. (2011). *Embedded formative assessment*. Bloomingdale: Solution Tree Press.
Wu, J. R. W., & Wu, R. Y. F. (2010). Relating the GEPT reading comprehension tests to the CEFR. *Studies in Language Testing, 33*, 204–224.

# Washback, Impact, and Consequences Revisited

Dina Tsagari and Liying Cheng

**Abstract**

Washback, impact, and consequences refer to the educational phenomenon when testing (often large-scale and high-stakes), specifically the uses of test scores and the decisions made based on those scores, influence those stakeholders associated with such testing. Washback, impact, and consequences are used in different fields of research, and these terms encompass different dimensions of the research undertaken. *Washback* is more frequently used to refer to the effects of tests on teaching and learning at the classroom level. *Impact* refers to the effects that a test may have on individuals, policies, or practices, within the classroom, the school, the educational system, or the society as a whole. Many language testers these days consider *washback* as a dimension of *impact*. The effects of testing on teaching and learning have been traditionally associated with test validity (*consequential validity*) where washback is considered as only one form of testing *consequences* that need to be weighted in evaluating validity. This chapter elaborates the origins and dimensions of these terms by presenting the major empirical studies conducted over the past 30 years. Considering the complexity of this educational phenomenon and increasing importance of the testing effects in education and beyond, the authors present the challenges facing such research and point out the directions that future research in this area could embrace.

**Keywords**

Washback (backwash) • Impact • Consequences • Teaching • Learning • Consequences of tests • Effects of tests - Large scale high-stakes testing • Tests •

D. Tsagari (✉)
Department of English Studies, University of Cyprus, Nicosia, Cyprus
e-mail: dinatsa@ucy.ac.cy

L. Cheng
Faculty of Education, Queen's University, Kingston, ON, Canada
e-mail: liying.cheng@queensu.ca

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_24

Examinations • Teaching and learning • Validity (consequential validity) • Measurement-driven instruction • Test-curriculum alignment • Ethics • Fairness

## Contents



## Introduction

It is accepted nowadays that "testing has become big business" (Spolsky 2008, p. 297) and that it plays a powerful role in education, politics, and society in general (McNamara and Shohamy 2008). High-stakes large-scale testing in particular "is never a neutral process and always has consequences" for its stakeholders (Stobart 2003, p. 140), intended or unintended, and positive or negative.

In the long and substantial amount of research conducted in general education, researchers refer to the phenomenon as *measurement-driven instruction* (Popham 1987), *test-curriculum alignment* (Shepard 1990), and *consequences* (Cizek 2001). By contrast, in language education, test consequences are a relatively new concept since the late 1980s. The two terms commonly used in the field are *impact* and *washback*. Wall (1997) defines *impact* as "any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole". She also points out that "*washback* (also known as *backwash*) is sometimes used as a synonym of impact, but it is more frequently used to refer to the effects of tests on teaching and learning" (p. 291) at the classroom level. Many language testers these days consider *washback* as a dimension of *impact* (e.g., Hamp-Lyons 1997).

Primarily, the effects of testing on teaching and learning have been associated with test validity (*consequential validity*) where Messick refers to washback as "only one form of testing consequences that need to be weighted in evaluating validity" (Messick 1996, p. 243). He stresses the need for the examination of two threats to test validity, *construct under-representation* and *construct-irrelevant variance*, to decide the possible consequences that a test can have on teaching and learning. Bachman (2005) proposes a framework with a set of principles and procedures for linking test scores and score-based inferences to test use and the consequences of test use. Other contemporary validity theories (Chalhoub-Deville 2015; Chapelle et al. 2010; Kane 2013, 2016) adopting the various argument-based models have established grounds for the inclusion of test consequences and uses within validation

studies. These theories require the systematic collection of validity evidence at each validation stage and from multiple stakeholder perspectives to better justify the use of test scores in pedagogical and policy practices.

In addition, the effects of testing on teaching and learning are increasingly discussed from the point of view of critical language testing, including ethics and fairness in language testing, all of which are expressions of social concern. For example, Shohamy (2001) points out the political uses and abuses of language tests and called for examining the hidden agendas of the testing industry and of high-stakes tests. Kunnan (2000, 2004) discusses the role of tests as instruments of social policy and control. He also draws on research in ethics to link validity and consequences and created a test fairness framework. Hamp-Lyons (1997) argues for an encompassing ethics framework to examine the consequences of testing on language learning at the classroom, as well as the educational, social, and political levels. All of the above has led to the creation of a Code of Ethics for the International Language Testing Association (see Davies 2008).

## Early Developments

The work of Alderson and Wall (1993) marked a significant development in shaping the constructs of washback studies for the field of language testing. The authors explored the potential positive and negative relationship between testing, teaching, and learning, and questioned whether washback could be a property of test validity. They consequently proposed 15 hypotheses (revisited and refined in Alderson and Hamp-Lyons 1996) regarding the potential influence of language testing on various aspects of language teaching and learning, which thus directed washback studies for years to come. The study of Wall and Alderson (1993) was the first empirical research published in the field of language testing. It investigated the nature of washback of a newly introduced national English examination in Sri Lanka by observing what was happening inside the classroom.

A review of the early literature, as pointed out by Cheng (2008), indicates at least two major types of washback studies. First there are those relating to traditional, multiple-choice, large-scale standardized tests; these are perceived to have had mainly negative influences on the quality of teaching and learning. Secondly there are those studies where a specific test or examination has been modified and improved upon (e.g., assessment with more communicative tasks: see Cheng 2005) in order to exert a positive influence on teaching and learning.

In 1996, a special issue in *Language Testing* published a series of articles that further explored the nature of washback and empirically investigated the relationship between testing, teaching, and learning. In this volume, Messick (1996) suggested building on validity considerations through test design in order to promote positive washback and to avoid construct under-representation and construct-irrelevant variance. Although Messick did not specify how researchers could go about studying washback through test design validation, he pointed out that test washback could be

associated with test property. He thus offered a coherent argument to investigate the factors in testing that are related to factors in teaching and learning. Bailey (1996, p. 268), however, argued that any test, whether good or bad in terms of validity, can have either negative or positive washback "to the extent that it promotes or impedes the accomplishment of educational goals held by learners and/or program personnel". Her argument indicated that washback effects (positive or negative) might differ for different groups of stakeholders. Finally, Wall (1996) stressed the difficulties in finding explanations of how tests exert influence on teaching, and turned to innovation theory to offer "insights into why attempts to introduce change in the classroom are often not as effective as their designers hoped they would be" (p. 334).

Three empirical research studies are also reported in the same special issue. Alderson and Hamp-Lyons (1996) found that the Test of English as a Foreign Language (TOEFL) affects both what and how teachers teach, but the effect is not the same in degree or kind from teacher to teacher. Watanabe (1996) found that teacher factors, including personal beliefs, past education, and academic background, seemed to be more important in determining the methodology a teacher employs rather than the university entrance examination in Japan. Shohamy et al. (1996) added that the degree of impact of a test is often influenced by several other factors: the status of the subject matter tested, the nature of the test (low or high stakes), the uses to which the test scores are put and that the washback effect may change over time.

In summary, testing may be only one of those factors that "affect how innovations [through testing] succeed or fail and that influence teacher (and pupil) behaviors" (Wall and Alderson 1993, p. 68). The special issue editors of the volume also call for the "need for co-ordinated research into washback and other forms of impact, and for a theory which will guide testers so that they have the best chance of influencing teaching, education and society in a positive way" (Alderson and Wall 1996, p. 240). Indeed the years since the 1996 special issue in *Language Testing* have seen a flurry of publications ranging from collections of empirical studies, doctoral theses, and research projects investigating different tests within different teaching and learning contexts. These will be presented in the following sections.

## Major Contributions

Two edited volumes that have become the cornerstone collection of washback studies and initial attempts to capture the essence of washback have been published in the 2000s. The first one was the publication of Cheng and Watanabe, with Curtis's *Washback in Language Testing: Research Context and Methods* (2004). Through its compilation of washback studies, the book responded to the question "what does washback look like?" – a step further from the question "does washback exist?" posed by Alderson and Wall (1993). In its first section, the volume highlights the concept and nature of washback by providing a historical review of the phenomenon (Cheng and Curtis 2004); the second section showcases a range of studies on various aspects of teaching and learning conducted in many parts of the world, e.g.,

Australia, China, Hong Kong, Israel, Japan, New Zealand, the UK, and the USA. The book has contributed to our understanding of washback and impact of language tests in that we can no longer take for granted that where there is a test, there is a direct effect.

The second publication was the special issue dedicated to investigating washback in language testing and assessment published in *Assessment in Education: Principles, Policy and Practice* in 2007. The editors, Rea-Dickins and Scott (2007), brought together papers from equally varied contexts that looked at washback areas such as the consequences of large-scale school tests on different groups of learners; the effects of a statutory national assessment curriculum on primary school learners; a specific writing task of a high-stakes test on secondary school students; three different program types on the development of students' writing skills; and the impact of the International English Language Testing System (IELTS) preparation classes on improving student scores on the writing sub-test of the test. The papers included also problematize on the selection of appropriate methodologies for researching washback and on how language tests are used as a mechanism in the manipulation of language education policies and policy control. The volume has added to our understanding of washback as being context-specific, unstable, and difficult to predict, and makes a call for greater dialogue between language and education researchers.

Several major doctoral studies have also made a substantial contribution to the understanding of the complexity of washback and offered methodological implications for washback studies over the years. For example, the longitudinal study by Wall (2005) documents research examining one of the widely held beliefs that change can be created in an education system by introducing or by re-designing high-stakes examinations. Wall analyzed the effects of a national examination in English as a Foreign Language in Sri Lanka that was meant to serve as a lever for change. Her study illustrated how the intended outcome was altered by factors in the exam itself, as well as the characteristics of the educational setting, the teachers, and the learners. Her study, located in the interface of examination impact and innovation in education, provided guidelines for the consideration of educators who continue to believe in the potential of examinations to affect curriculum change.

Through a large-scale, three-phase study using multiple methods to explore the multivariate nature of washback, Cheng (2005) investigated the impact of the Hong Kong Certificate of Education in English (HKCEE) on the classroom teaching and learning of English in Hong Kong secondary schools where the examination is used as the change agent. The washback effect of this public examination change was observed initially at the macro level, including various stakeholders within the local educational context, and subsequently at the micro level, including aspects of teachers' and learners' attitudes, teaching contents, and classroom interaction. The findings indicated that the washback effect of the new examination on classroom teaching was limited despite expectations to the contrary. Her study supports the findings of Wall and Alderson (1993), i.e., that the change of the examination can inform what teachers teach, but not how they teach.

Green (2007) used a variety of data collection methods and analytical techniques to explore the complex relationship between teaching and learning processes and their outcomes. Green evaluated the role of *IELTS* in English for Academic Purposes (EAP) particularly in relation to the length of time and amount of language support needed by learners to meet minimally acceptable standards for English-medium tertiary study. This piece of research is of relevance to a range of interested parties concerned with the development of EAP writing skills.

In her study, Tsagari (2009) explored the washback of First Certificate in English (FCE, offered by the formerly known Cambridge ESOL) on the teaching and learning that takes place in intermediate level EFL classes in Greece. The study followed a mixed-method design to data collection and analysis, e.g., interviews, teaching, exam-preparation materials and student diaries. The findings showed that many other factors beyond the test, the teachers or students (e.g., publishers/authors, the school, and the educational context) need to be taken into account when studying the washback effect of a high-stakes exam to explain why washback takes the form it does in a given context. The study led to a comprehensive model of exam washback and suggestions for teachers, teacher trainers, students, material and test developers, as well as future researchers in the area.

Many more doctoral level washback studies have been conducted over the years, which add to our understanding of the complex nature of washback and impact. These studies have been conducted in various contexts investigating the influence of testing on teachers and teaching, textbooks, learners and learning, attitudes toward testing, classroom conditions, recourse provision and management practices within the school, the status of the subject being tested in the curriculum, feedback mechanisms between the testing agency and the school, and the general social and political context. The studies have also focused on the influence of national examinations in countries such as Brazil, Canada, China, Egypt, Hong Kong, Iran, Israel, Japan, Spain, Taiwan, and the UK. Others have also looked at worldwide English testing such as IELTS, TOEFL, TOEIC, Cambridge Young Learners English test series, and the Michigan Examination for Certificate of Competency. For a review of washback and impact doctoral studies, see Cheng and Fox (2013) and Cheng et al. (2015).

The research output of the above studies shows that washback is a highly complex phenomenon due to the fact that it is an interactive multidirectional process involving a constant interplay of varying degrees of complexity among the different washback components and participants. Also the above studies have shown that simply changing the contents or methods of an examination will not necessarily bring about direct and desirable changes in teaching and learning. Rather various factors within educational contexts are involved in engineering desirable washback, e.g., test factors (test methods, test contents, skills tested, purpose(s) of the test), prestige factors (stakes of the test, status of the test), personal factors (teachers' educational backgrounds and their beliefs), micro-context factors (the school/university setting), and macro-context factors (the specific society in which the tests are used) (Cheng and Curtis 2004). However, questions remain about the nature of

factors and stakeholders involved, the interaction between them and the conditions under which beneficial washback is most likely to be generated.

## Work in Progress

The interest in test washback and impact continues to grow, as evidenced in major conference presentations such as Language Testing Research Colloquium (LTRC) and publications in journals such as *Language Testing* and *Language Assessment Quarterly* and edited volumes or monographs.

Several important projects have been commissioned by major testing agencies and have increasingly played a major role in producing clusters of washback and impact studies. These studies are conducted in many countries around the world on the same test and tend to be large-scale and multi-faceted. They offer important recommendations for the improvement of the tests under study and directions for future research. For instance, findings of funded research studies, such as those conducted by Cambridge English language assessments, report on the impact of examinations at the micro (teaching and learning) and at macro levels (employability, schools, parents, and decision makers) in countries such as Cyprus, Greece, Japan, Romania and Spain. Long-term research on IELTS has been implemented through the Cambridge ESOL Research & Validation Group. Most of these studies are also collaborative in nature, which indicates the importance of working with local experts, and employ mixed-method designs. The results are published online via Research Notes, RN (http://www.cambridgeenglish.org/research-and-validation/published-research/research-notes/) and other publications (Hawkey 2006). This flurry of research has resulted in a richer and more informative picture of washback and impact, and a more in-depth understanding of the current state of English language teaching, learning, and assessment within the particular contexts.

Educational Testing Services (ETS) has also funded a series of studies examining the impact of the TOEFL test. For example, the TOEFL Impact Study in Central and Eastern Europe (Wall and Horák 2006, 2008, 2011) investigated whether the new TOEFL iBT contributed to changes in teaching and learning after its introduction. This study involved three research stages: Phase 1: a "baseline study" described the type of teaching and learning taking place in commercial language teaching operations before details of the test were released about the content and format of the new test; Phase 2: a "transition study" traced the reactions of teachers and teaching institutions to the news that was released about the TOEFL iBT and the arrangements made for new preparation courses; Phase 3 investigated whether textbooks published accurately reflected the new test and what use teachers make of them in the classroom. Data was collected via computer-mediated communication with informants providing responses to the activities in their classrooms and institutions and reactions to tasks, which had been designed to probe their understanding of the new test construct and format (see also Hamp-Lyons and Brown 2007; Tsagari 2012).

Funded by the Social Sciences and Humanities Research Council of Canada (SSHRC), Cheng and colleagues' large collaborative study on *Test Preparation: Does It Enhance Test Performance and English Language Proficiency* (http://educ.queensu.ca/test-prep) is a multiphase and multiyear investigation into the relationship of test preparation and test performance (Cheng and Doe 2013). This study is conducted in partnership with major test agencies and various stakeholders: test developers, teachers, students, test preparation center administrators and staff, and university admissions officers. The researchers conducted case studies of test preparation courses in Australia, Canada, China and Iran linking students' test preparation practices to their test performance (Ma and Cheng 2016; Saif et al. 2015). The study findings will provide test-designers and test users with empirical evidence regarding the predominant phenomenon of test preparation and the validity of test scores.

## Problems and Difficulties

Although there have been increasing numbers of empirical washback and impact studies conducted since the late 1980s, researchers in the field of language education continue to wrestle with the nature of washback, and to research ways to induce positive and reduce the negative washback and impact of language tests. As indicated above, washback is one dimension of the consequences of the testing on classroom teaching and learning, and impact studies include broader effects of testing (as defined in Wall 1997). However, both assume a causal relationship between testing, teaching, and learning which has not been established up to now. Most of the washback and impact empirical studies have only established an exploratory relationship. In many cases, we cannot be confident that certain aspects of teaching and learning perceptions and behaviors are the direct causal effects of testing. They could well be within certain contexts, but this relationship has not yet been fully disentangled.

Furthermore, apart from the studies on IELTS, TOEFL, and large collaborative studies, e.g., Cheng and colleagues' study on test preparation, where a worldwide test influences teachers and learners across countries and educational contexts, the majority of the empirical studies focuses on the effects of one single test, within one educational context using research instruments designed specifically for that particular study. The strength of such studies is that they have investigated factors that affect the *intensity of washback* (Cheng and Curtis 2004). In fact, many of the factors related with the influence of testing on teaching and learning illustrated in Wall (2000) have been empirically studied. However, not only does little overlap exist among the studies regarding what factors affect washback, but little overlap also exists in researcher reports of the negative and positive aspects of washback (Brown 1997). In addition, there does not seem to be an overall agreement on which factors affect the intensity of washback and which factors promote positive or negative washback. This is a challenging feature of washback and impact studies, since

researchers set out to investigate a very complex relationship (causal or exploratory) among testing, and teaching and learning.

This complexity causes problems and difficulties in washback and impact research, which in turn challenges any researcher who wishes to conduct, is conducting, or has conducted such studies. In many ways, the nature of such washback and impact study requires subtle, refined, and sophisticated research skills in disentangling this relationship. Researchers need to understand the specificity, intensity, length, intentionality, and value of test washback/impact and how (or where and when) to observe the salient aspects of teaching and learning that are potentially influenced by the test. They also need to identify their own bias, analyze the particular test and its context, and produce the predications of what washback and impact looks like prior to the design and conduct of the study (see also Watanabe 2004). Washback and impact studies are, by definition, studies of program evaluation, which require researchers not only to understand but also to make a value judgment about the local educational context as well as the larger social, political, and economic factors governing teaching and learning in relation to a test/examination or a testing system. Researchers need to acquire both the breadth and depth of necessary research skills to avoid research based on investigating random factors of teaching and learning, which may or may not have a direct relationship with testing.

## Future Directions

It is clear that the future direction of washback and impact studies to investigate the consequences of language testing need to be multi-phase, multi-method, and longitudinal in nature. Washback and impact of testing take time to evolve, therefore longitudinal studies are essential with repeated observations (and measures) of the classroom teaching, including teachers and students as well as policy, curriculum, and assessment documents. Also, researchers need to have very good knowledge and understanding of the test they investigate, work collaboratively with the test developers and be well-immersed in the educational system they investigate, interacting with a wide range of stakeholders. In addition, researchers should pay attention to the seasonality of the phenomenon, i.e., the timing of researchers' observations may influence what we discover about washback (Bailey 1996; Cheng 2005) avoiding potential bias. Examples like the IELTS impact studies and the impact studies on TOEFL iBT across different countries and continents over a few years have a great deal to contribute to our understanding of this complex phenomenon. Studies of a single test within an individual context by a single researcher can still offer valuable insights for that particular context; however, it would be best if groups of researchers could work collaboratively and cooperatively to carry out a series of studies around the same test within the same or across educational contexts. The findings of such research could then be cross-referenced to portray a more accurate picture of the effects of the test avoiding the "blind men and elephant" syndrome. Research studies need also to move from the micro-level of the classroom (washback) to the macro-level of society (impact), to analyze the social factors that lead to assessment

practices in the first place, and to explain why assessment practices (large-scale testing) are valued more than others. Such studies also need to link the (mis-)uses of test scores with what happens inside and outside the confines of the classrooms.

In addition, the methodology (and the methods) used to conduct washback studies need to be further refined. For example, researchers need to vary their methods including mixed method explanatory, exploratory, and concurrent design. Also, more sophisticated data collection and analysis methods, e.g., those linking directly with test-takers' characteristics, perceptions of assessments, learning processes, and their learning outcomes (test performance) need to be employed beyond classroom observations and survey methods (interviews and questionnaires) (e.g., Xie and Andrews 2013). Building on the increasing numbers of studies carried out on the same test or within the same educational context, future researchers can replicate or refine instruments and analysis procedures, which was not possible in the past. The replication would allow researchers to build on what we have learned theoretically, conceptually, and methodologically over the years and further our understanding of this phenomenon.

While it would be useful to continue to study the effects of tests on broad aspects of teaching, it is essential to turn our attention to investigate the effects on students and their learning as they receive the most direct impact of testing. In other words, what has not been focused on in previous studies is the direct influence of testing on learners (e.g., their perceptions, their strategy use, motivation, anxiety, and affect), on their learning processes (e.g., what and how they learn, or how they perform on a test including test-taking), and learning outcomes (test scores or other outcome measures). Based on these investigations, it is also important to use the results to do in-depth observations of students. For example, it is crucial to study students' understanding of the test constructs especially in the public examination domains where test-related information may not be directly accessible to students. In addition, it is important to study factors that are likely to be shaped by the learning and wider societal context. Other than students' perceptions of a test, it is also important to examine how they obtain such knowledge. This type of research can directly link the consequences of testing with test validity. It would be also worthwhile to look at the test taker population more closely, e.g., the educational characteristics (in terms of learning and testing) of the students. We know by now that high-stakes testing like IELTS or TOEFL influences students. However, is the impact of the test different on students learning English in one country than in another where the educational tradition (beliefs and values) are different? Without a thorough understanding of where these students come from and the characteristics they bring to their learning and testing, it is unlikely that we can fully understand the nature of test washback and impact.

Research also needs to be directed towards the relationship between high stakes large-scale testing and classroom-based teacher-led formative assessment (Tsagari and Banerjee 2014). Research in this area can better inform teachers for their curriculum planning and instruction and can better support student learning, making ongoing teacher involvement a part of test development and validation process

(Froetscher 2016). Another fruitful area of research is the investigation of "language assessment literacy" – LAL (Fulcher 2012, p. 125) of high-stakes test users, e.g., teachers (Vogt and Tsagari 2014), university admissions officers (O' Loughlin 2013), test writers, and professional language testers (Harding and Kremmel 2016). It is important to understand the degrees of LAL needed for different stakeholder groups in high-stakes test contexts as this can induce positive consequences from tests. This is an exciting research venue that will be able to attest to the urgent and largely unrecognized need in many high-stakes educational and policy-making contexts for increasing teacher development opportunities (Taylor 2013).

An additional area that lacks empirical research is washback on stakeholders outside the immediate confines of the classroom, e.g., parents, who tend to be neglected, but take important instructional decisions about learners outside of their school time. In addition, given that assessment is located in the social context, empirical studies of indirect participants – such as public media, language accreditation systems, employers and policy makers – and their perceptions and understanding of high-stakes tests and use of test scores are needed, as these will add greater importance to the washback phenomenon and unveil different degrees of complexity.

Finally it remains controversial in educational assessment research whether and how consequences should be integrated in test validation or even whether they belong to test validation or not (Messick 1989; Moss 1998; Nichols and Williams 2009). A few studies in the field of language assessment have systematically investigated test consequences within a coherent validation framework to examine evidence for the purpose of evaluating the strength of the validity argument (including consequential validity) of a particular test in a given situation (Chapelle et al. 2010; Chalhoub-Deville 2015). In the end, washback and impact researchers need to fully analyze the test under study and understand its test use. Bachman (2005, p. 7) states that "the extensive research on validity and validation has tended to ignore test use, on the one hand, while discussions of test use and consequences have tended to ignore validity, on the other". It is, then, essential to establish the link between test validity and test consequences theoretically and empirically. It is imperative that washback and impact researchers work together with other language testing researchers, as well as educational policy makers and test agencies, to address the issue of validity, in particular, fairness and ethics of language tests.

## Cross-References


- ▶ Critical Language Testing
- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ Language Assessment Literacy
- ▶ Training in Language Assessment

## Related Articles in the Encyclopedia of Language and Education

Linda von Hoene: The Professional Development of Foreign Language Instructors in Postsecondary Education. In Volume: Second and Foreign Language Education

Bonny Norton, Ron David: Identity, Language Learning and Critical Pedagogies in Digital Times. In Volume: Language Awareness and Multilingualism

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

## References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A case study. *Language Testing, 13*, 280–297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

Alderson, J. C., & Wall, D. (1996). Editorial. *Language Testing, 13*, 239–240.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257–279.

Brown, J. D. (1997). Do tests washback on the language classroom? *The TESOLANZ Journal, 5*, 63–80.

Chalhoub-Deville, M. (2015). Validity theory: Reform policies, accountability, testing, and consequences. *Language Testing*. doi:10.1177/0265532215593312.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge: Cambridge University Press.

Cheng, L. (2008). Washback, impact and consequences. In N. H. Hornberger (Series Ed.), *Encyclopedia of language and education* (Language testing and assessment, 2nd ed., Vol. 7, pp. 349–364). New York: Springer.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3–18). Mahwah: Lawrence Erlbaum Associates.

Cheng, L., & Doe, C. (2013). "Test preparation: A double-edged sword", *IATEFL-TEASIG (International Association of Teachers of English as a Foreign Language's Testing, Evaluation and Assessment Special Interest Group). Newsletter, 54*, 19–20.

Cheng, L., & Fox, J. (2013). Review of doctoral research in language assessment in Canada (2006–2011). *Language Teaching, 46*, 518–544.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Mahwah: Lawrence Erlbaum Associates.

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*, 436–470.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 23*(3), 1–17.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy, N. H. Hornberger (Eds.), *Encyclopedia of language and education*. (Language testing and assessment, 3rd ed., Vol. 7, pp. 429–444). New York: Springer.

Froetscher, D. (2016). A new national exam: A case of washback. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 61–81). London: Continuum.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge, UK: Cambridge University Press.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295–303.

Hamp-Lyons, L., & Brown, A. (2007). *The effect of changes in the new TOEFL format on the teaching and learning of EFL/ESL: Stage 2 (2003–2005): Entering innovation.* Report submitted to the TOEFL research committee, Educational Testing Service.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 413–427). Berlin/New York: Muton De Gruyter.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Kane, T. M. (2013). Validating the interpretations and uses of test scores. *Educational Testing Service Journal of Educational Measurement, 50*(1), 1–73.

Kane, T. M. (2016). Explicating validity. *Assessment in Education: Principles Policy and Practice, 23*(2), 198–211.

Kunnan, A. J. (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium Orlando Florida*. Cambridge: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic, C. Weir, & S. Bolton (Eds.), *Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27–48). Cambridge: Cambridge University Press.

Ma, J., & Cheng, L. (2016). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth and significance. *TESL Canada Journal, 33*(1), 58–79 . http://www.teslcanadajournal.ca/index.php/tesl/article/view/1227.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89–95.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 243–256.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3–9.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappa, 68*, 679–682.

Rea-Dickens, P. & Scott, C.: (2007). Investigating washback in language testing and assessment' (Special issue). *Assessment in Education: Principles, Policy and Practice*, *14*(1), 1–7.

Saif, S., Cheng, L., & Rahimi, M. (2015). *High-stakes test preparation programs and learning outcomes: A context-specific study of learners' performance on IELTS*. Paper presented at the 37th annual language testing research colloquium. Toronto, 16–20 Mar 2015.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*, 15–22.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex: Longman.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*, 298–317.

Spolsky, B. (2008). Language testing at 25: Maturity and responsibility. *Language Testing, 25*(3), 297–305.

Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education: Principles, Policy and Practice, 16*, 139–140.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Tsagari, D. (2009). *The complexity of test washback: An empirical study*. Frankfurt am Main: Peter Lang GmbH.

Tsagari, D. (2012). *The influence of the Examination for the Certificate of Proficiency in English (ECPE) on Test Preparation materials*. Internal report sponsored by the SPAAN fellowship for studies in Second or Foreign Language Assessment, Cambridge Michigan Language Assessments (CaMLA), Ann Arbor.

Tsagari, D., & Banerjee, J. (2014). Language assessment in the educational context. In M. Bigelow & J. Ennser-Kananen (Eds.), *Handbook of educational linguistics* (pp. 339–352). New York: Routledge/Taylor & Francis Group.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*, 334–354.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (pp. 291–302). Dordrecht: Kluwer Academic Publications.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28*, 499–509.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge: Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41–69.

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe: Phase 1, The baseline study, TOEFL monograph series; MS-15*. Report number: RR-06-18, TOEFL-MS-34. Princeton: Educational Testing Service, http://www.ets.org/Media/Research/pdf/RR-2006-2018.pdf

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change*. TOEFL iBT research report. TOEFLiBT-05, https://www.ets.org/Media/Research/pdf/RR-08-37.pdf

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL® exam on teaching in a sample of countries in Europe: Phase 3, The role of the coursebook phase 4, describing change*. TOEFL iBT® research report TOEFL iBT-17, http://www.ets.org/Media/Research/pdf/RR-2011-2041.pdf

Watanabe, Y. (2004). Methodology in Washback Studies. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 19–36). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*, 318–333.

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation: Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49–70.

# Part IV

# Assessment in Society

# History of Language Testing

Bernard Spolsky

**Abstract**

Although some thirty years ago I saw language testing as made up of three successive periods, each an advance over the last, I am now less optimistic. Rather, I see it as developed as part of a long history of examinations, starting with the Imperial Chinese system and moving from the selection among an elite to an effort to control mass education systems. The result has been industrialization, so that testing has become big business, and political concern for accountability is threatening to swamp schools with tests. To handle evidence of inevitable uncertainty, psychometrics developed techniques to show statistical reliability, but efforts to demonstrate validity remain inconclusive, though construct validity and argument-based approaches focused on test use are suggesting promising leads. Testers have developed guidelines for ethical testing, but there is no enforcement. Computerization has raised new problems but not solved old ones. Language testers are open to the implications of language diversity, and some propose multilingual testing. But the power of the established systems continues.

**Keywords**

Uncertainty • Psychometrics • Industrialization • Ethics • Scales

## Contents



B. Spolsky (✉)
Bar-Ilan University, Ramat Gan, Israel
e-mail: bspolsky@gmail.com

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_32

## Introduction

The challenge in the title that I have been assigned is daunting; if only I still had the chutzpa that I had 30 years ago when I happily divided the history of language testing into three periods (Spolsky 1977)! Since then, having actually spent many months two decades later studying the history of the field (Spolsky 1995b), I am much less confident about my ability to describe its development and quite certain that any predictions that I may make have little chance of being correct: as the Talmud points out, now that prophecy is dead, only fools and babies venture to do it.

What I plan to do in this short paper is to sketch out a number of key events in the history of testing in general and language assessment in particular, using them as labels for clusters of approaches to assessing human abilities that continue to show up in one form or another under varying circumstances. Because it is not always clear that a new approach is derived from an older one, I will not argue that these add up to any record of progress in the field (this was the basic flaw in my 1977 paper, where I was attempting to argue that the current approach was the best). There were influences, no doubt, but each is best seen as an adaptation to contemporary concerns and possibilities. Nor will I take the position that so many critics do: *swā lengra, swā wyrs* – things go on getting worse, as the bad drives out the good.

In any history of testing, the Chinese Imperial examinations naturally come first, as this was the first state-wide effort to establish a testing system under centralized control. At times during the 2000-year history of the system, the Emperor himself saw the final papers. The aim of the examination system was to winnow out of a large pool of candidates the very best could be selected for government office as Mandarins, with subsequent major financial reward. The Chinese principle, as Lord Macaulay (1853) called it when arguing in the British Parliament for its adoption as a method of selecting cadets for the Indian Civil Service, involved a long and complex academic examination intended to test and rank a number of well-prepared candidates. It had no interest in evaluating or influencing an educational system, but its only concern was sorting and selecting a tiny and in theory independent elite. While the Imperial examinations faded away even before the last of the Emperors (Franke 1960), they had a major effect in establishing a very strong (and eventually unhealthy) respect for testing in China and other countries under Chinese influence, producing continued pressure on education in Japan and Korea as well as China.

This first major examination may well be contrasted with the medieval Treviso test, in which the students of the Northern Italian school were assessed at the end of the year by representatives of the city council which then paid the schoolmaster according to the success of his students (Ariès 1962; Madaus 1990). The important characteristics of this approach were its focus on the educational process, its use of the school curriculum as specification of content, and its interest in ranking pupils only in terms of their mastery of the curriculum. When in the sixteenth century the Jesuits brought the Chinese notion of examinations to Europe, they adapted it to the Treviso purpose of the control of the curriculum: in the classical Christian Schools of the seventeenth and eighteenth century (de la Salle 1720), pupils were tested at

regular intervals on their mastery of the curriculum, and their progress was determined by the success on the regular tests.

During the French Revolution, the religious schools were secularized, and their examination system was further modified to suit the needs of a strong central government. Napoleon (another Emperor) established academies in each department to be responsible for the administration of the centrally controlled examination, the core of which was the Bac (baccalaureate) which marks the culmination of secondary school study (Théry 1861). The system continues in effect, its emphasis on government control proving popular in many other countries as well.

The oral examination for lieutenants that Samuel Pepys introduced into the British Navy at the end of the seventeenth century (Tomalin 2003) had features of the Chinese Imperial Examination in that it wished to replace selection by patronage by merit and of the Treviso test in that its concern was not with ranking but with mastery of a defined body of knowledge. Carried out by senior captains, it had the authority of their experience but the unreliability of untrained judges.

In England in the nineteenth century, there were important new developments. Macaulay (1853, 1898) argued in Parliament for the examinations that were adopted initially for the Indian and later for the English Civil Service. They were partially modeled on examinations at Oxford and Cambridge, the goal of which was to rank a comparatively small number of well-prepared candidates; at Oxford, those who did well received first, second, or third class honors; and at Cambridge, the best student in the Tripos was called senior Wrangler. A similar model was adopted in nineteenth-century Prussia in the selection of magistrates (McClelland 1980). In England and the colonies, the popular esteem in which examinations were held made it possible toward the end of the century to widen the system: examinations were commonly conducted by school inspectors to assess the achievement of pupils in state elementary schools.[1]

In the early years of the twentieth century, there was an effort to extend this school-based testing by the use of the objective techniques being developed by the new field of psychometrics. In England, the interest was short lived (Burt 1921) but revived in 1944 with the eleven plus examination intended to decide what kind of secondary school a pupil should go to, but in the United States, where a number of interested parties trumpeted the mythical success of intelligence testing by the army in the First World War (Reed 1987; Yerkes 1921), the 1920s were marked by the proliferation of tests and of testing companies ready to sell them to schools. The American or objective test (initially true-false but later increasingly multiple choice) came to be seen as the ideal instrument for most assessment purposes. Occasionally adopted elsewhere, it met resistance in much of the world until the full forces of globalization after the Second World War led to its rapid proliferation and current virtual universalization. It is now associated with a movement for accountability in education, with many systems accepting the Treviso principle that teachers should be paid according to their pupils' test results.

[1] I found a report of my father's score on a 1910 report when he was 11.

Language testing grew up against this background. The early tests often included components intended to assess language competence. Where the emphasis was on literacy skills, the method was generally to require composition or translation (and later, as these do not lend themselves to objective testing, comprehension which can be tested with multiple-choice items). When oral skills were considered important (in sixteenth century Cambridge University where students were still expected to know Latin well enough to use it or in testing spoken language ability), there was oral testing, but it proved particularly difficult to adjust to the requirements of large-scale objective testing (Spolsky 1990). In special cases, however, such as the comparatively small-scale elite testing for the US diplomatic service (Wilds 1975), there were reasons and resources to develop a comparatively standard method of assessing spoken ability, and there continue to be attempts to do it more cheaply using computers or telephones.

This, as they used to say in the old continuous movie houses, is where we came in. For many language testers of my generation, the history of our field starts in the 1960s, the beginning of the large-scale industrialization and centralization of language testing that has come to be based in Princeton and Cambridge. My 1977 view of this point in time was a progression from a "traditional" examination (consisting of written translation, composition, comprehension, and grammar) through a psychometrically driven testing of structural linguistic items that had formed the basis of such tests as the Michigan Lado test of English (Lado 1951). In the next stage, there appeared a new trend that combined John Carroll's argument for integrative testing (Carroll 1961) with the experience of the FSI oral examination, all modified in the light of Cooper's argument for adding sociolinguistic aspects (Cooper 1968). To this, one would now add the attraction of computerization of the process, both test administration and (still controversial but growing) and scoring.

Looking back over the half-century during which language assessment has developed into an identifiable academic field as well as a major industry, there are several trends which are worth identifying. One, particularly relevant to the academic field but with strong influence on practical test development, has been the effort to overcome what was recognized a hundred years ago as the unavoidable uncertainty of examinations (Edgeworth 1888). The field of psychometrics has been struggling ever since to find ways of making tests reliable and valid. Once statistical methods of establishing reliability were found, replacing single individual measures like essays with large numbers of objective items lending themselves to appropriate statistical treatment, testers could argue that their test was reliable: in other words, that it would have much the same result when repeated on other occasions or with other candidates. More difficult has been agreement on the validity, essentially the meaning rather than the stability of the result. In the early days of language testing, tests were considered valid if they correlated well with other tests. The justification for a new test was a fairly high correlation (say .8, which meant that two-thirds of the variation had been accounted for) with some existing test it was meant to replace.

The last 50 years have seen much more robust and intelligent efforts to establish validity. There have been rare efforts to validate the predictive power of a test: the early versions of the IELTS were validated by asking university tutors whether the

results agreed with their own judgment of English knowledge of foreign students who had taken it (Criper and Davies 1988). For many years, ETS suggested that each institution should carry out its own validation study of the TOEFL results of students they admitted; very few did this, however. Another approach has been construct validation, an effort to build a theoretical model of the ability being measured and then to determine that test items could reasonably be assumed to measure the various described aspects of the construct (Weir 2005; Weir et al. 2013). The notion of validity was greatly broadened by the work of Lyle Bachman (1990, 2000) who applied to language testing the extended definition of validity proposed by Messick (1980, 1989) which included focus on impact. The pursuit of validity continues: Bachman has now found ways to integrate the social implications and the use of language test results into the model, and Weir's work is central to test revision at Cambridge.

A second major trend has been the complications involved in what Carroll called integrative testing, the assessment of samples of language (written or spoken) produced by the candidates and hardly susceptible to objective measurement by requiring human judgment. Language testers in the 1930s and 1940s wanted to test these performances but were challenged by the psychometric difficulties of establishing reliability on the one hand and in determining which factors led to individual judgments on the other. The use of such tests were also practically difficult: in the 1940s, the Cambridge English testers had to use Post Office engineers to record samples of oral tests in order to try to train new judges; in 1961, the ETS representative easily dissuaded the TOEFL planners from including a writing sample because of the expense of air mailing examination booklets to the US; in the mammoth Chinese English Test, oral testing by two examiners was restricted to 100,000 or so of the 6 million candidates who took the test each time it was given.

But there continued to be pressure. During the brief period of the Army Specialized Training Program (Iglehart 1997; Spolsky 1995a), Kaulfers (1944) planned but never implemented a scale for oral testing of the soldiers in the program. When later the Assistant Secretary of State insisted that American diplomats be tested for their language proficiency, oral ability could not be left out, and the Foreign Service Institute, with advice from John Carroll, developed a scale and began a system of testing using two or three judges. A number of years later, information about the test was made available to academic language testers (Wilds 1975), and it became the model for such testing in other government agencies and later for foreign language testing in the United States and elsewhere.

We thus had two major trends: a pursuit of reliability that provided backing for the development of industrialized objective tests, and a market-driven demand for more or less valid measurement of productive proficiency. The first of these led to the development of the large-scale industrial test. While most early language tests were the work of individual language teachers and testers, starting in the 1920s in the USA, they were quickly taken up by small psychological testing corporations most of which were in due course swallowed by publishers, only to be taken over in due course by large international conglomerates (Clark and Davidson 1993). The exception was Educational Testing Service, born as the testing arm of the College Board

and provided with permanence and independence by being set up in New Jersey as a nonprofit corporation licensed by the New York Board of Regents. Their major moneymaker in language testing was TOEFL, the Test of English as a Foreign Language, and was set up originally as an independent body but brought under ETS control in a series of brilliant political maneuvers (Spolsky 1995b). For most of its 40 years at ETS, TOEFL was a prime example of an industrial test, open to market forces rather than to changing theory. Both the Test of Spoken English and the Test of Written English were in response to demands from test users rather than independent innovation.

In England, the process was similar. English language tests were produced by a number of testing centers, generally affiliated with universities, but by the 1970s the University of Cambridge Local Examination Syndicate had clearly taken the lead. The market forces in the period after the Second World War with growing demand for English language teaching and testing persuaded the Syndics to take the field seriously, and they learnt their lesson well from an attempt to compete with TOEFL by claiming to be comparable in reliability and validity (Bachman et al. 1989). In particular, they made sure that their English testing division remained independent and was able to make use of its growing profits to carry out the research needed for constant improvement of their tests (Weir and Milanovic 2003).

Perhaps the most significant development from the nonindustrial oral testing process was the growth of the importance of the scale. For Thorndike in the early days of testing, a scale consisted of a number of prejudged and carefully ranked examples of the product being scaled: handwriting or essays (Thorndike 1910). For the Foreign Service Institute, a scale was a verbal protocol describing as accurately as possible the particular characteristics of a language performance of a defined level or stage of learning (Jones 1979). Such scales worked as mnemonics for trained judges to remind them of the consensus they had reached in training exercises and in previous experience. They raise all sorts of intriguing theoretical problems: they assume, for instance, that language proficiency is scalable rather than a set of partially related abilities in performing various language functions. As a result, they need to be accepted by consensus rather than validated in practice or theory. One of the most elaborate developments of the language scale is the Common European Framework (Council of Europe 2001), which in fact comprises a large number of different scales for various kinds of language knowledge and functional ability. Given the convenience of scales however for practical use, the tendency has been to attempt to reduce the Framework to a relatively simple scale equivalent to the US Interagency Roundtable scale.

Besides what may be called the psychometric, industrial, and scaling trends, an important development in late twentieth-century language testing was the broadening of the content to include sociolinguistically influenced aspects of language. It was Cooper (1968) who pointed out the need to include social context, and with the development of communicative language teaching and lip service at least to ethnography of speaking (Hymes 1967, 1974), language testing has broadened from the academic testing of the standard written version of a language to allow for assessment of control of other varieties in various social and functional situations. One

inevitable conclusion has been the realization that tests need to find some way to achieve authenticity and to measure the ability to perform in situations not unlike the real world.

There has been another important adjustment of language testing to its social context, and this has been the attempt to define and even judge the social context of test use and the ethicality of the test. The concern over social and educational effects of tests was not new: Latham (1877) complained that examinations at the end of the nineteenth century were leading to narrowing of educational goals. British testers like Cyril Burt were conscious of the social influence of tests, hoping they would permit more intelligent children from the lower socioeconomic classes to become upwardly mobile. Similarly, Lemann (1999) argues that Chauncey's main motive in developing the SAT was to bring a wider socioeconomic population into top East Coast colleges and universities and so into the national leadership.[2] In language testing, too, we have come to be concerned about social implications. Edelsky et al. (1983) suggested that unwillingness of disadvantaged pupils to play the testing games led to misconceptions about their language proficiency. Spolsky (1981, 1984) argued for the need to take an ethical view of the effects of a language test. Shohamy (1992, 1994, 1997, 2001) took this concern further, detailing the power of language tests for social control. Most recently, McNamara (2005) and other scholars have described the use of language tests outside the school system to verify identity of people claiming ethnicity or asylum and to filter or block immigrants. The effect has been felt within the profession (Hamp-Lyons 1997), and over the past few years, professional language testing associations and groups have been working to develop codes of ethics and of professional practice (see http://www.iltaonline.com/code.pdf).

There have been some new developments since I wrote the first version of this paper. One has been an important recognition of multilingualism especially in cities, sometimes labeled as a modern effect of globalization producing super-diversity (Cadier and Mar-Molinero 2012; De Angelis 2014; Romaine 2008; Shohamy 2011). Shohamy has argued for multilingual tests. Another has been the recognition of language diversity, as pointed out by Kachru (1986) and leading to questioning of testing of standard British or American English.[3] Much of this is focused on the notion of English as a lingua franca (Seidlhofer 2011), which has led to a tendency to demote the native speaker as goal.

Having almost run out of space, I am spared the task of spending very much time on future trends. Were I still a young optimist, I would suggest that everything will continue to get better. Industrial tests will become more human and less powerful; only valid tests will be used (after careful validation) for major career decisions; simple unidimensional scales and scores will be replaced by complex profiles showing the wide range of plurilingual proficiency of anyone tested; tests will not

[2]It didn't work, but it did allow women and Jews and later Asians into Ivy league schools.

[3]When Robert Lado showed his tests to the UCLES staff in 1959, it was the Americanisms rather than the objective items that shocked them.

be misused. But from my current perspective, I am much more skeptical. I see the industrial test makers working industriously to computerize their tests and sell them wherever possible; I read online discussions in which writers painfully and hesitatingly try to rerun debates about cloze tests that were closed decades ago; I see multidimensional profiles being reduced to uniform scales; I see one whole establishment refusing to recognize that the highly educated native-speaker-like ambassador is not the only top of the language tree; I see countless school and university systems continuing to interpret more or less randomly awarded scores as though they were meaningful. At the same time, I expect to continue to see good research into the nature of language proficiency and the continued demonstration of possible ways to assess its relevance to defined social purposes.

In other words, more of the same. An idea, I notice, for which I am quoted in the introduction to the testing volume of the first edition of this encyclopedia (Clapham 1997).

## Cross-References

- ▶ Assessing Meaning
- ▶ Critical Language Testing
- ▶ Culture and Language Assessment
- ▶ Ethics, Professionalism, Rights, and Codes

## Related Articles in the Encyclopedia of Language and Education

Sandra Lee McKay: Sociolinguistics and Language Education. In Volume: Second and Foreign Language Education

Kate Menken, Ofelia Garcia: Language Policy in Classrooms and Schools. In Volume: Language Policy and Political Issues in Education

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

Bernard Spolsky: Language Policy in Education: History, Theory, Praxis. In Volume: Language Policy and Political Issues in Education

## References

Ariès, P. (1962). *L'enfant et la vie familiale sous l'Ancien Régime*. Paris: Plon.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we test counts. *Language Testing, 17*(1), 1–42.

Bachman, Lyle F., Davidson, Fred, Ryan, K, & Choi, I. (1989). An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study: final report: University of Cambridge Local Examinations Syndicate.

Burt, C. L. (1921). *Mental and scholastic tests*. London: London County Council.

Cadier, L., & Mar-Molinero, C. (2012). Language policies and linguistic super-diversity in cointemporary urban societies: The case of the City of Southampton UK. *Current Issues in Language Planning, 13*(3), 149–165.
Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.
Clapham, C. (1997). Introduction. In C. Clapham & D. Corson (Eds.), *Language testing and assessment*. Dordrecht: Kluwer.
Clark, J. L. D., & Davidson, F. (1993). Language-learning research: Cottage industry or consolidated enterprise. In A. O. Hadley (Ed.), *Research in language learning: Principles, process, and prospects* (pp. 254–278). Lincolnwood: National Textbook.
Cooper, R. L. (1968). An elaborated language testing model. *Language Learning, 18*(Special issue No. 7), 57–72.
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Criper, C., & Davies, A. (1988). ELTS Validation Project Report: The British Council and the University of Cambridge Local Examinations Syndicate.
De Angelis, G. (2014). A multilingual approach to analysing standardized test results: immigrant primary school children and the role of languages spoken in a bi-/multilingual community. *Intercultural Education, 25*(1), 14–28. doi:10.1080/14675986.2014.883167.
de la Salle, J.-B. (1720). *Conduite des Ecoles chrétiennes*. Avignon: C. Chastanier.
Edelsky, C., Altwerger, B., Flores, B., Hudelson, S., & Jilbert, K. (1983). Semilingualism and language deficit. *Applied Linguistics, 4*, 1–22.
Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599–635.
Franke, W. (1960). *The reform and abolition of the traditional Chinese examination system*. Cambridge, MA: Harvard University Center for East Asian Studies.
Hamp-Lyons, L. (1997). Ethics in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Language testing and assessment, Vol. 7, pp. 323–334). Dordrecht: Kluwer.
Hymes, D. (1967). Models of the interaction of language and social setting. *Journal of Social Issues, 23*(2), 8–38.
Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.
Iglehart, F. N. (1997). *The short life of the ASTP*. Baltimore: American Literary Press.
Jones, R. L. (1979). The oral interview of the Foreign Service Institute. In B. Spolsky (Ed.), *Some major tests* (pp. 104–115). Washington, DC: Center for Applied Linguistics.
Kachru, B. B. (1986). *The alchemy of English: The spread, functions and models of non-native Englishes*. Oxford: Pergamon Institute of English.
Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *Modern Language Journal, 28*(2), 136–150.
Lado, R. (1951). *English language tests for foreign students*. Ann Arbor: George Wahr.
Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Deighton, Bell and Company.
Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Strauss and Giroux.
Macaulay, T. B. (1853). *Speeches, parliamentary and miscellaneous*. London: Henry Vizetelly.
Macaulay, T. B. M. B. (1898). *The Works of Lord Macaulay*. London, Longmans, Green.
Madaus, G. P. (1990). *Testing as a social technology*. Boston: Boston College.
McClelland, C. (1980). *State, society and university in Germany, 1700–1914*. Cambridge: Cambridge University Press.
McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351–370.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Reed, J. (1987). Robert M. Yerkes and the mental testing movement. In M. M. Sokal (Ed.), *Psychological testing and American society 1890–1930* (pp. 75–95). New Brunswick: Rutgers University Press.
Romaine, S. (2008). Language rights, human development and linguistic diversity in a globalizing world. In P. v. Sterkenburg (Ed.), *Unity and Diversity of Languages* (pp. 85–96). Amsterdam: John Benjamins Publishing Company.
Seidlhofer, B. (2011). *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.
Shohamy, E. (1992). *The power of tests: a study of the impact of language tests on teaching and learning.* Paper presented at the Language Testing Research Colloquium, Vancouver.
Shohamy, E. (1994). The use of language tests for power and control. In J. E. Alatis (Ed.), *Georgetown University Round Table on language and linguistics* (pp. 57–72). Washington, DC: Georgetown University Press.
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340–349.
Shohamy, E. (2001). *The Power of tests: A critical perspective of the uses of language tests*. London: Longman.
Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal, 95*(3), 418–429.
Spolsky, B. (1977). Language testing: Art or science. In G. Nickel (Ed.), *Proceedings of the Fourth International Congress of Applied Linguistics* (Vol. 3, pp. 7–28). Stuttgart: Hochschulverlag.
Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5–30). Frankfurt am Main: Verlag Peter D. Lang.
Spolsky, B. (1984). The uses of language tests: An ethical envoi. In C. Rivera (Ed.), *Placement procedures in bilingual education: education and policy issues* (pp. 3–7). Clevedon/Avon: Multilingual Matters.
Spolsky, B. (1990). Oral examinations: An historical note. *Language Testing, 7*(2), 158–173.
Spolsky, B. (1995a). Behind the ASTP myth. In G. H. Richins & R. K. Belnap (Eds.), *Selected Papers from the Proceedings of the Twentieth Annual Symposium of the Deseret Language and Linguistics Society, 3–4 March 1994* (pp. 119–124). Provo Utah: Deseret Language and Linguistics Society.
Spolsky, B. (1995b). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.
Théry, A. F. (1861). *Histoire de l'éducation en France, depuis le Cinquième Siècle jusqu'à nos jours* (2nd ed.). Paris: E. Magdeleine Dezobry.
Thorndike, E. L. (1910). Handwriting. *Teachers College Record, 11*(2), 83–175.
Tomalin, C. (2003). *Samuel Pepys: The unequalled self*. London: Viking.
Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
Weir, C. J., Vidaković, I., & Galaci, E. D. (2013). *Measured constructs:A history of Cambridge English language examinations 1913–2012*. Cambridge, UK: Cambridge University Press.
Weir, C., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. Cambridge: Cambridge University Press.
Wilds, C. (1975). The oral interview test. In B. Spolsky & R. L. Jones (Eds.), *Testing language proficiency* (pp. 29–37). Washington, DC: Center for Applied Linguistics.
Yerkes, R. M. (Ed.). (1921). *Psychological examining in the United States Army.* Washington, DC: Government Printing Office.

# High-Stakes Tests as De Facto Language Education Policies

Kate Menken

**Abstract**

The practice of high-stakes testing, using a single test score to determine major decisions, has become commonplace in school systems around the world. Tests now hold extremely important consequences for students, teachers, schools, and entire school systems. Because the stakes of current tests are so high, they guide a wide range of educational choices, including curricula, materials, pedagogy, teacher preparation, educational programming, and language medium of instruction. In this way, high-stakes tests result in de facto language education policies. This chapter examines the relationship between high-stakes testing practices and language education policies to show how testing becomes de facto language policy in schools.

The chapter begins with a brief exploration of the history of the standardized testing movement. It then presents recent empirical research that has investigated high-stakes testing from a language education policy perspective in order to deepen understandings of how high-stakes standardized tests become de facto language policies in implementation, as schools respond to the exacting pressures of testing. The chapter documents the detrimental impact of monolingual testing, focusing on the formal education of children and how students and the education they receive is affected by recent testing practices. It then explores the potential for multilingual assessment informed by recent translanguaging theory to address these issues.

K. Menken (✉)
Department of Linguistics, Queens College, City University of New York, Flushing, NY, USA

Research Institute for the Study of Language in an Urban Society, Graduate Center, City University of New York, New York, NY, USA
e-mail: kmenken@gc.cuny.edu; kmenken@qc.cuny.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_25

**Keywords**
Language education policy • High-stakes testing • Standardized testing • Translanguaging

## Contents



## Introduction

There has been an increase in the use of high-stakes standardized testing over the past 20 years in school systems around the world, whereby important consequences are attached to a single test score, often as part of standards-based educational reform efforts (also known as outcomes-based reforms). A test becomes high stakes when a single test score is used as the main or sole factor in determining significant educational decisions. Internationally, single test scores carry high-stakes consequences for individual students, as they are used to determine achievement, learning level, grade promotion, grade retention, attainment of a diploma, and university admission. Raising the stakes even further, student performance results on these tests are also being used by governments to evaluate teachers, schools, and entire school systems as a means to hold educators and educational systems responsible for student learning. In this chapter, I examine the relationship between high-stakes testing practices and language education policies to show how testing becomes de facto language policy in schools.

In order to do so, it is first necessary to clarify how language education policy is defined (for a more detailed discussion of language policy, see "Language Policy and Political Issues in Education," of this Encyclopedia). Language policy refers to "formal and informal decisions about language use, which includes laws, regulations, and statutes, as well as practice" (de Jong 2011, p. 256), and involves language practices, beliefs, and management (Spolsky 2004). Language education policies determine which language(s) are taught and used as medium of instruction, how they are taught, and how linguistic diversity is negotiated in schools (Shohamy 2006; Spolsky 2004; Tollefson 2012). Thus while language policies are concerned with the decisions that people make about languages and their use in society, language education policies refer to carrying out such decisions in educational contexts (Shohamy 2006). School language policies are of pressing concern because they can determine language maintenance or oppression, with long-lasting implications for speakers of a given language and their communities.

High-stakes testing thus carries major implications for language education and emergent bilinguals (students who speak a language other than the national or dominant

language at home and are learning it in school). In addition to either rewarding or barring an individual student from future opportunities, such tests are used in schools to guide a wide range of educational choices, including curricula, textbooks and materials, pedagogy, teacher preparation, programming, and language medium of instruction. The higher the stakes of a test, the greater impact it has on the education that students receive. In this way, high-stakes tests act as de facto language education policies.

Yet rarely do policymakers or test developers consider the language policy byproducts of standardized tests. Instead, the language policies embedded within high-stakes exams are typically implicit rather than explicit, though extremely powerful in shaping changes at the classroom level (Menken 2008; Shohamy 2001). Overlooking the language policy implications of high-stakes testing has proven particularly harmful for emergent bilinguals, their teachers, and their schools. As research in language policy has become more attentive to policy implementation in schools (taking a "bottom-up" as well as "top-down" perspective), recent empirical research has studied high-stakes testing from a language education policy perspective. This line of research is presented in this chapter, as it examines how high-stakes standardized tests become de facto language policies in implementation, as schools respond to the exacting pressures of testing.

## Early Developments

This review begins by exploring the history of the standardized testing movement, with particular attention to the implementation of such tests with speakers of minoritized languages, such as immigrants and indigenous populations worldwide. Today's standardized tests have historical roots in the mental measurement movement that focused on intelligence quotient (IQ) testing. Spolsky (1995) reports that the whole testing movement initially flourished in the United States and spread globally after World War I. The development of intelligence tests, and IQ testing in particular, coincided with a rapid increase in immigration to the United States at the turn of the twentieth century (Hakuta 1986). Alfred Binet is credited with creating the first IQ test in 1904 at the request of the French government, for identifying children to be placed into special education programs. In 1917, after being translated into English, IQ tests were used by H. H. Goddard to test immigrants arriving in the United States via Ellis Island. Of 30 adult Jews tested, 25 were found to be "feeble minded" (Hakuta 1986, p. 19).

Carl Brigham, one of the founders of the testing movement, administered IQ tests in English to two million World War I draftees in the United States and analyzed why test takers born in the United States or in the United States for 20 years or more outperformed recent immigrants (Spolsky 1995). Basing his findings on national origin, race, ethnicity, and English literacy, he found that blacks were inferior to whites. Brigham categorized Europeans as "Nordics, Alpines, and Mediterranean races" and found that Alpine and Mediterranean races were inferior to the Nordic race (Hakuta 1986; Wiley and Wright 2004). Hakuta (1986) critiques these findings

with regard to immigrants, for failing to acknowledge how language proficiency is directly tied to exam performance.

The political use of IQ testing was not limited to immigrants, as the tests were used as a sorting mechanism in education for all students; immigrants as well as other minorities have historically been particularly vulnerable to high-stakes decisions made on the basis of test scores. IQ tests justified racial segregation of US schools in the twentieth century, and test scores resulted in the hierarchical ranking of students within schools of that era (Mensh and Mensh 1991). In addition, the findings Brigham reported in his book, *A Study of American Intelligence,* influenced Congress to pass an act restricting immigration by "non-Nordics." As Wiley and Wright (2004) summarize:

> English literacy became a gatekeeping tool to bar unwanted immigrants from entering the United States when nativists began clamoring for restrictions. Simultaneously, literacy requirements barred African Americans at the polls. . .Thus, the so-called scientific testing movement of the early 20th century was intertwined with racism and linguicism at a time when the push for expanded uses of restrictive English-literacy requirements coincided with the period of record immigration. (pp. 158–159)

From the beginning, testing has been exploited as a means to exert power, authority, and control. Scientifically proven to be neutral and impartial, tests very effectively sort, select, and punish (Shohamy and Menken 2015). The sections that follow explore the intersection between testing and language education policy.

## Major Contributions

Shohamy (2001) was the first author to argue that language testing is in actuality de facto language policy, particularly when high stakes are attached. As she wrote in the introduction to her book, *The Power of Tests*:

> Professor Bernard Spolsky and I were asked to propose a new language policy for Israel. Given my background and interest in language testing, I again learned about the power of tests as it became clear to me that the 'language testing policy' was the *de facto* 'language policy'. Further, no policy change can take place without a change in testing policy as the testing policy becomes the *de facto* language policy. (Shohamy 2001, p. xiii)

In Israel, a new test of Arabic as a foreign language was introduced that was intended to raise the prestige of the language among Hebrew speakers. After several years, Shohamy (2001) found that the test influenced teaching, learning, and curricula to such an extent that teaching and testing had essentially become synonymous. Even so, it had not successfully raised the status of the Arabic language in Israel. These cases show how policymakers use tests to create de facto policies that will promote their agendas and communicate their priorities, a top-down practice which Shohamy (2001) characterizes as unethical, undemocratic, and unbeneficial to the test taker.

More recently, Shohamy (2006) highlights the following as the three major language policy implications of testing: determining prestige and status of languages, standardizing and perpetuating language correctness, and suppressing language diversity (p. 95). For example, the use of the Test of English as a Foreign Language (TOEFL) internationally to determine school or university entrance contributes to the high status and prestige of English as a global language.

Uniformity of approach and content is a common result of high-stakes testing, both in foreign language and second language education. One example is the various language proficiency rating scales provided by US government agencies, such as the Foreign Service Institute, the Defense Language Institute, and the Peace Corps, and the widespread use of the Common European Framework of Reference (CEFR) in Europe. Such scales establish set stages of language learning, as though learning a language follows a prescribed and controlled linear order (Shohamy and Menken 2015). The CEFR is particularly powerful in education, where it has become a prescriptive sequence of how and what learners learn – a problematic interpretation that has been critiqued by Fulcher (2004) and Shohamy (2006), among others.

Language assessment research on what is termed testing "washback" offers further evidence for tests as de facto language policies in education. In early research in this area, Messick (1996, p. 241) defined washback as "the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning." The argument that washback can be either positive or negative is reinforced more recently by Fulcher (2010) as well as Cheng and Curtis (2012), who argue further that washback should be purposeful in order to encourage positive effects. The findings in this area of research, however, seem to indicate that the effects are more often negative and unintended.

One major critique of high-stakes standardized testing in the washback literature is that curricula and instruction are narrowed as a consequence, such that the material on the test drives what is taught and "teaching to the test" becomes commonplace. This is reported as far back as 1802, when a new exam was introduced at Oxford and criticized because it resulted in a student's education becoming narrowed down to only the subjects being assessed (Simon 1974 as cited in Wall 1997). Hayes and Read (2004) demonstrate how changes to the English language testing system in New Zealand show washback effects, with teachers and students narrowly focused on test tasks rather than on academic language proficiency in the broader sense.

In research on the National Matriculation English Test (NMET) in China, Qi (2005) showed how the exam did not have the positive effects that were intended, as teaching of linguistics knowledge rather than communicative competence continued to be emphasized and the elements of language taught remained limited to the skills tested. As the author stated:

> When crucial decisions are made on the basis of test results and when one's interests are seriously affected, who can afford not to teach to or study for the test? This is especially true when the sole measure of the success of the educational process being evaluated is the test scores. . . (Qi 2005, p. 163)

The higher the stakes of an exam, the more likely it is that the test will shape teaching, learning, and language policy. As the author explains, the notion of using high-stakes testing to positively change teaching is innately problematic, due to the pressure it creates to teach to the test.

Cheng (2004) studied changes to language testing in Hong Kong and concluded that the washback effects were negative because they resulted in drilling what was required by the exam. By definition, this restricted the scope of what students learned, and drilling or rote memorization activities overshadowed possibilities for authentic language use. From these studies, it seems that the possibilities for positive washback effects from high-stakes testing are limited by the nature of the exams themselves.

The more recent literature in this area has focused on efforts to promote what is termed "positive washback" or "intended washback" (see, for instance, Cheng and Curtis 2012; Muñoz and Álvarez 2010). For example, in their research on English as a foreign language instruction in Colombia, Muñoz and Álvarez (2010) argue for giving teachers "constant guidance and support over time" in order to generate positive washback. While washback research explores the ways that testing changes teaching and learning, little attention is paid in the washback literature to the effects of testing on the lives of educators and students in schools. This differs from research on testing from a language policy perspective, which goes beyond the impact of testing on language education to also consider the social justice implications when a test acts as de facto language policy.

## Work in Progress

Recent research highlights the problems associated with monolingual testing for emergent bilinguals and language education policy and the potential for multilingual assessment to address these issues. Research conducted in the United States shows how monolingual testing serves to marginalize emergent bilinguals and their language practices, limit their future opportunities, and encourage monolingual language policies in schools. US federal education policy entitled *No Child Left Behind* (NCLB) was passed into law by Congress in 2001.[1] Under the law, emergent bilinguals are to be tested annually in English language proficiency and academic content, with failure resulting in high-stakes consequences for schools and school systems (e.g., federal sanctions such as school closure or loss of funding) as well as for students (e.g., grade promotion and graduation; Menken 2008). NCLB is found to encourage instruction in English only, particularly due to its accountability

[1]It is worth noting that as this chapter goes to press, NCLB has been replaced by a new federal law called the *Every Student Succeeds Act* (passed into law in December 2015). While this new law allows for greater state autonomy on testing, all states still have the same assessment systems in place as they had under NCLB. To date there has been no reduction of high-stakes testing in US schools, and the full impact of this new legislation remains to be seen.

mandates, as emergent bilinguals and their schools must prepare for high-stakes tests in English and are disproportionately likely to fail and be penalized (Menken 2010; Wiley and Wright 2004). A number of studies (Evans and Hornberger 2005; Menken 2008; Menken and Solorza 2014) show how this policy of monolingual testing raises the status of English while it suppresses non-tested languages and has directly contributed to the elimination of bilingual education programs.

For instance, Crawford (2004) reports that NCLB testing policies undermine bilingual education programs when the tests are provided in English only. In a dual language bilingual education program in Montgomery County, Maryland, instructional time was balanced equally between English and Spanish. However, the school district became concerned with poor reading scores by emergent bilinguals on high-stakes exams and mandated two-and-a-half-hour blocks of English phonics each day. This increased the amount of daily English instruction, which disrupted the bilingual program's equal instructional time in each language.

Wiley and Wright (2004) note how the word "bilingual" was removed from NCLB. They critique the inclusion of emergent bilinguals into tests in English when they have not had sufficient time to acquire the language, highlighting the similarity between current high-stakes tests and literacy and intelligence tests administered during the early twentieth century to bar immigrants from entering the United States. They argue that although NCLB does not prohibit bilingual programs, it does encourage English-only approaches (Wiley and Wright 2004).

Menken's (2008) qualitative research in New York City examines the requirement that all students pass a set of state exams in order to graduate from high school, as part of the state's accountability system under NCLB. Due to the high stakes of these assessments, tests have become de facto language policy in city schools where Menken (2008) argues that they shape what content is taught in school, how it is taught, by whom it is taught, and in what language(s) it is taught. As she writes:

> The tests themselves leave the task of interpretation to teachers and schools, who decipher their demands and use them to create a complex and wide array of school-level language policies. While it is tempting to assume that top-down policy will simply be unidirectional in implementation, and that if *No Child Left Behind* implicitly promotes English then English will always be favored in instruction. In actuality, however, this assumption is overly simplistic; while most schools in this sample indeed increased the amount of English instruction students receive to improve their test performance, one school and certain teachers were found doing exactly the opposite, and instead increased native language instruction as a test preparation strategy. (Menken 2008, p. 11)

Particularly concerning is Menken's (2008) finding that the testing requirements in New York to comply with NCLB have resulted in increased dropout rates and decreased graduation rates for emergent bilinguals, only one-third of whom successfully graduate from high school each year.

These findings are recurrent nationally, where emergent bilinguals disproportionately fail high-stakes standardized tests and are placed into low-track remedial education programs, denied grade promotion, retained in grade, and/or leave school (Gándara and Contreras 2009; Vasquez Heilig 2011; Valenzuela 2005). Vasquez

Heilig (2011) and McNeil (2005) report a narrowing of school curricula and large amounts of instructional time devoted to test preparation in Texas classrooms, particularly for emergent bilinguals. Untested subjects such as science and social studies are being abandoned at the elementary level, because tests of English and math are the areas required under NCLB.

Empirical research by Palmer and Lynch (2008) offers further support for the argument that the link between testing and language policy is not always unidirectional. While these authors highlight the tension between *mono*lingual testing and *bi*lingual education, they show that within bilingual programs, the language favored in instruction is the same language in which students will be tested. Based on qualitative research conducted in six elementary schools in Texas, a state where high-stakes tests of certain subjects can be taken in English or Spanish, Palmer and Lynch (2008) show how it is testing that drives language education policy decisions. As these authors write:

> We argue that children who test in Spanish will be taught in Spanish, with little attention to the transition process until the testing pressures are lifted; children who test in English will be taught in English, with little attention to the support in their primary language that may determine their ability to succeed on a test in their second language. (Palmer and Lynch 2008, p. 217)

As described here, teachers "teach to the test" by matching the language of instruction to the language of the tests.

Although New York is also a state where test translations are available, research by Menken and Solorza (2014) shows a causal link between the testing and accountability policies of NCLB and the loss of bilingual education programs in New York City schools. Menken and Solorza (2014) conducted qualitative research in ten New York City schools that had recently eliminated their bilingual education programs in order to determine the factors that drove that decision. They note how emergent bilinguals in city schools in 2001 were evenly divided between bilingual education and English as a second language (ESL) programs, but since then the proportion of emergent bilinguals enrolled in bilingual education has decreased to just 22%, while ESL enrollment has increased to over 76%. Menken and Solorza (2014) show how test-based accountability is problematic for schools serving emergent bilinguals, as these students are disproportionately likely to be labeled low performing due to the impact of language proficiency on test performance across all subjects. This places school administrators under enormous pressure to improve the performance of their students so that their schools can be deemed successful under NCLB, and many respond to this pressure by adopting English-only language policies for their schools. As the authors write:

> Bilingual education programs are immediately blamed for the poor performance of emergent bilinguals on high-stakes tests and other measures of accountability. . .Principals in our sample turn to language programming changes for emergent bilinguals as a way to provide the "quick fix" their schools needs to immediately meet the federal and local accountability requirements. (Menken and Solorza 2014, p. 108)

Although the availability of test translations might create an opening for bilingual educators to teach in a language other than English, the New York case shows how this alone is not enough. In New York, test translations fail to curtail the widespread elimination of bilingual education programs in the wake of high-stakes standardized testing.

Moving away from test translations, which are monolingual in that they require students to answer test items in one language or the other (not both), multilingual assessment is a promising possibility to address some of the problems associated with monolingual testing described above. Multilingual assessment is grounded in very recent research in applied linguistics about translanguaging that pushes back against rigid language separation or what Cummins (2005, p. 588) terms the "two solitudes" or García (2009, p. 70) refers to as "monolingualism times two." This research clarifies how the languages of bilinguals do not work in isolation from one another, but rather are deeply interconnected. García (2009) offers the term translanguaging to describe bilingual language practices and capture the flexible and complex ways they language in order to make meaning.

Shohamy (2011) proposes the use of multilingual testing to better match the actual language practices of emergent bilinguals. Empirical research by Shohamy (2011) as well as by Rea-Dickins et al. (2011) suggests that the use of multilingual assessments significantly contributes to higher scores on academic tasks and more accurately reflects the knowledge of test takers. Shohamy and Menken (2015) argue that a translanguaging approach to multilingual assessment should drive future research and practices in language testing. As they write:

> [W]hat we term here "dynamic assessment" offers affordances for emergent bilinguals to use their entire linguistic repertoire flexibly and creatively to process and produce language for academic purposes through various procedures such as mediation and displaying test questions simultaneously in two languages. Rather than suppress students' home language practices, given the power and prestige of language tests, dynamic multilingual assessments not only offer more accurate information about students and improve their outcomes, but also serve to raise the prestige of students' home languages in schools and society. (Shohamy and Menken 2015, p. 265)

That said, very little research to date has been conducted about the actual development and implementation of multilingual assessments. As such, this is a promising area for future research.

## Future Directions

As described above, language policies are created by high-stakes testing at every level of educational systems around the world in ad hoc, uncoordinated, and often competing ways. More often than not, this is done implicitly, with the language policy implications of tests rarely being discussed openly or explained from the outset. Yet tests wield enormous power over the lives of students and educators and

shape how testing policy is exercised in schools and societies. For example, they affect the instruction and educational experiences of students in school and also determine students' futures.

Whether done explicitly or implicitly, the findings from the studies cited here show that the effects of high-stakes tests as de facto language policy are often unintended. Research about the intersection between testing and language policy is recent, and overall there is very little research on this critical topic. Yet major decisions are being made in school systems every day based on test scores. At the school level, curriculum and teaching are narrowed to the material on the tests, and certain languages are privileged over others in education. Moreover, tests can offer a justification for the perpetuation of societal inequities in schools, a trend from the past being repeated in schools today.

In light of these complex issues and limited research on testing as language education policy, there is a need for further research in this area. The following are possibilities for future directions:

- More research is needed that explores how testing shapes and affects language education policies and offers guidance for the use of assessment to inform educational practices and instruction in positive ways.
- It would be useful to learn if the use of multiple measures of student achievement (e.g., the use of portfolios, an array of samples of student work, grades, classroom performance, and teacher recommendations) for high-stakes decision-making would have the same impact on language medium of instruction, language standardization, and language status that the use of a single test score has had.
- Likewise, research on the development of clear and cohesive school-wide language policies in individual schools would be valuable to learn if schools that have developed their own language policies and established a strong vision for language education are as greatly impacted by the pressures of high-stakes testing.
- Research is needed on how assessment practices can be informed by the newer translanguaging paradigm, such as through multilingual assessments, and the impact of doing so.
- Many countries have adopted language policies that are multilingual, especially in recent years, yet testing practices have not followed suit. Research is needed to better understand the possibilities for aligned testing practices to multilingual language policies.
- The focus of this chapter has been on testing in formal educational contexts for children, yet testing serves many different purposes in language teaching and learning, for instance, in the education of adults, in tertiary education, in higher education, for professional certification, for determining citizenship, and so on. Research is needed on the interplay between testing practices and language education policies in these areas as well, so that their impact can also be better understood and mediated.

## Cross-References


- ▶ Critical Language Testing
- ▶ Dynamic Assessment
- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ History of Language Testing
- ▶ Language Assessment in Higher Education
- ▶ Washback, Impact, and Consequences Revisited


## Related Articles in the Encyclopedia of Language and Education

Jasone Cenoz: Translanguaging as a Pedagogical Tool in Multilingual Education. In Volume: Language Awareness and Multilingualism

Ofelia García, Angel Lin: Translanguaging in Bilingual Education. In Volume: Bilingual and Multilingual Education

Bernard Spolsky: Investigating Language Education Policy. In Volume: Research Methods in Language and Education

Bernard Spolsky: Language Policy in Education: History, Theory, Praxis. In Volume: Language Policy and Political Issues in Education

## References


Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng & Y. Watanabe (Eds.) with A. Curtis (Ed.), *Washback in language testing: Research contexts and methods* (pp. 147–170). Mahwah: Lawrence Earlbaum Associates.

Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, B. O'Sullivan, P. Davidson, & S. Stoynoff (Eds.), *Cambridge guide to second language assessment* (pp. 89–95). Cambridge: Cambridge University Press.

Crawford, J. (2004). *No child left behind: Misguided approach to school accountability for English language learners*. Paper for the Forum on ideas to improve the NCLB accountability provisions for students with disabilities and English language learners. Washington, DC: Center on Education Policy & National Association for Bilingual Education.

Cummins, J. (2005). A proposal for action: Strategies for recognizing heritage language competence as a learning resource within the mainstream classroom. *Modern Language Journal, 89*(4), 585–592.

de Jong, E. (2011). *Foundations for multilingualism in education: From principles to practice*. Philadelphia: Caslon.

Evans, B., & Hornberger, N. (2005). No child left behind: Repealing and unpeeling federal language education policy in the United States. *Language Policy, 4*, 87–106.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly, 1*(4), 253–266.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education/Routledge.

Gándara, P., & Contreras, F. (2009). *The Latino education crisis: The consequences of failed educational policies*. Cambridge, MA: Harvard University Press.

García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Malden: Wiley-Blackwell.

Hakuta, K. (1986). *The mirror of language: The debate on bilingualism*. New York: Basic Books.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng & Y. Watanabe (Eds.) with A. Curtis (Ed.), *Washback in language testing: Research contexts and methods* (pp. 97–112). Mahwah: Lawrence Earlbaum Associates.

McNeil, L. (2005). Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (Ed.), *Leaving children behind: How "Texas-style" accountability fails Latino youth* (pp. 57–112). Albany: State University of New York Press.

Menken, K. (2008). *English learners left behind: Standardized testing as language policy.* Bristol: Multilingual Matters.

Menken, K. (2010). *No child left behind* and English language learners: The challenges and consequences of high-stakes testing. *Theory Into Practice, 49*(2), 121–128.

Menken, K., & Solorza, C. (2014). No child left bilingual: Accountability and the elimination of bilingual education programs in New York City schools. *Educational Policy, 28*(1), 96–125.

Mensh, E., & Mensh, H. (1991). *The IQ mythology: Class, race, gender, and inequality.* Carbondale: Southern Illinois University Press.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256.

Muñoz, A., & Álvarez, M. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27*(1), 33–49.

Palmer, D., & Lynch, A. (2008). A bilingual education for a monolingual test? The pressure to prepare for TAKS and its influence on choices for language of instruction in Texas elementary bilingual classrooms. *Language Policy, 7*(3), 217–235.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173.

Rea-Dickins, P., Khamis, Z., & Olivero, F. (2011). Does English-medium instruction and examining lead to social and economic advantage? Promises and threats: A Sub-Saharan case study. In E. Erling & P. Seargeant (Eds.), *English and international development*. Bristol: Multilingual Matters.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman/Pearson Education.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. London: Routledge.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal, 95*(3), 418–429.

Shohamy, E., & Menken, K. (2015). Language assessment: Past to present misuses and future possibilities. In W. Wright, S. Boun, & O. García (Eds.), *Handbook of bilingual and multilingual education* (pp. 253–269). Hoboken: Wiley-Blackwell.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Spolsky, B. (2004). *Language policy.* Cambridge: Cambridge University Press.

Tollefson, J. (Ed.). (2012). *Language policies in education: Critical issues*. New York: Routledge.

Valenzuela, A. (Ed.). (2005). *Leaving children behind: How "Texas-style" accountability fails Latino youth*. Albany: State University of New York Press.

Vasquez Heilig, J. (2011). Understanding the interaction between high-stakes graduation tests and English language learners. *Teachers College Record, 113*(12), 2633–2669.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Language testing and assessment. Encyclopedia of language and education* (Vol. 7, pp. 291–302). Dordrecht: Kluwer.

Wiley, T., & Wright, W. (2004). Against the undertow: Language-minority education policy and politics in the "age of accountability". *Educational Policy, 18*(1), 142–168.

# Ethics, Professionalism, Rights, and Codes

Alan Davies

**Abstract**

The present chapter reviews aspects of ethics and professionalism in language testing and assessment and considers notions of rights and the use of codes as a way of linking both aspects. Against the traditional professions of law and medicine, language testing's claims to professionalism are not strong. But what it can do is to publish its commitment to ethics by means of a Code of Ethics. This provides for accountability both to members of the profession and to its stakeholders. This drive to accountability, to make its principles and practices explicit, explains the emphasis given in the language testing literature to the role of standards, both as goals and as the criteria for evaluating language testing procedures. It also explains the concern in the profession to uphold individual rights, especially those of test-takers. The review accepts that both professionalism and Codes of Ethics can be used improperly for face-saving ends and raises the question of how far issues to do with ethics, professionalism, rights and codes can be subsumed under the overall concepts of reliability and validity.

**Keywords**

Profession • Standards • Ethics • Codes • Rights • Validity

## Contents



A. Davies (✉)
University of Edinburgh, Edinburgh, UK

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_27

## Introduction

Before discussing the specific characteristics of professionalism in language testing and assessment, various definitions of a profession should be considered, allowing us to examine the claims of language testing for professional status in relation to general constructs of professionalism. This will enable us to understand language testing's decades-long efforts to professionalize itself in a broader context, and the devices it has used for achieving this goal.

## Early Development: Professions

Max Weber (1948) contrasted professions with bureaucracy, seeing in professions the paradigm form of collegiate activity in which rational power is based on representative democracy and leaders in principle are first among equals. Fullwinder provides the following criteria for a profession:

- It is a performance for the public good.
- It contains special knowledge and training.
- It deals mainly with people who for different reasons are especially vulnerable and dependent in their relationship to the practice of the professional (1996, p. 73).

Such criteria are readily applicable to the traditional professions of law and medicine, which explains Fullwinder's further comment that what distinguishes a profession from, say, a business, is its primary concern with the public good, since "that doctors and lawyers do not exploit . . . vulnerability, but help persons overcome serious threats to their health and rights constitute the great public good of the two professions" (1996, pp. 73, 74). And he suggests that whether or not an activity meets the criteria for a profession may be determined by completing the following schema:

> The profession of . . . serves the . . . needs of persons. (: 74)

Webster's Ninth *New Collegiate Dictionary of the English Language* (1994 edition) defines a profession as a calling. This primary definition "a calling" alerts us to

the derivation of profession: "professing, to profess: to commit oneself < profiteri, to declare publicly, to own or confess freely, to give evidence and thus to avow, in particular to declare oneself to be something (a friend, a philosopher, a physician, a teacher) entailing a pledge of capacity to fulfil the undertaking" (Siggins 1996, p. 56).

Siggins makes much of the early monastic influence:

> The profession of religion was the technical term for conversion to the monastic life and its vows . . . when universities appeared in a resurgent Renaissance Europe, it was first of all the teachers of sacred theology who were called 'professors' (p. 57) . . . the transition from cloister to university, however, laid the groundwork for the emergence of these disciplines (law, medicine) as proud, autonomous and eventually secular orders of society. (pp. 63, 64)

Marshall (1994, p. 419) defines a profession as "a form of work organization, a type of work orientation and a highly effective process of interest group control." Such an organization requires:

- A central regulatory body to ensure the standards of performance of individual members
- A code of conduct
- Careful management of knowledge in relation to members' expertise
- Control of entry numbers

There is a more sceptical view of the professions, querying their concern for the public good and seeing them as interest groups set up so as to exercise control over clients by means of socially constructed problems and thereby exert power. Ivan Illich (1987) saw the professions as totally self-interested and hypocritical. They created new needs among the general population and then made the public totally dependent. This approach treats professional ethics as an ideology rather than as an orientation necessarily adhered to or meaningful in practice. Marshall (1994) contends that, in such a setting, entry and knowledge controls function as a form of status exclusion for privileged and remunerative employment. And, somewhat ironically, while trade unions, that parallel (and very different) form of work sodality, become more professional in practice and orientation through, for example, job-entry controls, so the professions become more unionate, permitting, for example, collective bargaining and embracing industrial conflict.

In recent years many work-related activities have sought to describe themselves as professional. The reason for what has been called *the professionalization of everything* is, no doubt, greater public demand for accountability and widespread desire to emulate the status accorded to the law and medicine. The two professions of medicine and law, sometimes termed the noble professions, are revered as models of professionalism by those in other forms of employment, from estate agents and hairdressers to accountancy and language testing. They, like medicine and the law, demonstrate to their members and to the world that they are professions by publishing a Code of Ethics. The Codes claim professionalism, that is, faithfulness to the rules and articles of the occupation, but the higher the degree of professionalism

required of members, the stronger and more enforceable the code. Professions which are state licensed can enforce their codes through sanctions, such as dismissal from the profession, which are not available to weaker professions. Again, medicine and the law are canonical examples.

A Code (of Ethics and/or Practice) is one of the devices which provide for accountability by its apparent openness, thereby permitting the profession to publish its concern for the common good. Such codes set out the principles the profession binds itself to maintaining. Skene (1996) proposes that there are two types of code: the first type is intended to maintain standards of practice within the profession and to protect the community. The provisions of this first type of code may be prescriptive (and duty oriented) or aspirational (and virtue related). The second type is intended to protect the interests (especially the financial ones) of the profession and of its members, by including rules new members must accede to and requiring that only fully qualified people may be admitted to the profession, that members must be loyal to one another, and that they should not compete unfairly with one another.

## Work in Progress: Standards

The current drive for accountability may explain the frequent references everywhere to "standards" and therefore suggest to us that the standard concept is new and original. It is not. The search for standards has a long tradition, often under different names, the most common probably being norms, but there are other familiar terms too such as rules and conventions. What they all indicate is that there are social goals and that there are agreed ways of reaching towards those goals.

Brindley places standards under the broad heading of outcome statements: these, he considers, can refer to standards themselves and to benchmarks, attainment targets, bandscales, profiles, and competencies, all of which "are broadly speaking, standards of performance against which learners' progress and achievement can be compared" (1998, p. 48). Elder argues that within institutions, standards have more authority since they can be used as nonnegotiable goals (Elder 2000a).

In language assessment, standards have two senses. I note them here and then discuss each in turn:

1. The skills and/or knowledge required in order to achieve mastery and the proficiency levels leading to mastery, along with the measures that operationalize these skills and/or knowledge and the grades indicative of mastery at each level
2. The full set of procedures followed by test constructors which provide evidence to stakeholders that the test/assessment/examination/evaluation is serious and can be trusted, demonstrating, often through a Code of Ethics, that the test constructors are operating professionally

The two senses are also sometimes combined.

In the first sense, standards are the goal, the level of performance required or explained, thus "the standard required for entry to the university is an A in English";

"English standards are rising" (Davies et al. 1999, p. 185). Stakeholders, of course, rightly wish to know what is meant by such statements, how they are arrived at, and what is the evidence for making them. For this, there are three requirements: description, measurement, and reporting. There needs to be a description of the standard or level, an explicit statement of the measure that will indicate that the level has or has not been reached and a means of reporting that decision through grades, scores, impressions, profiles, and so on.

Description, measure, and report, these three stages are essential, although there may be blurring of stages 2 and 3, such that the report is included within the measure. Where classical objective tests such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) differ from the scale approaches of the Inter-Agency Round Table (IAR) and the American Council on the Teaching of Foreign Languages (ACTFL), the International Second Language Proficiency Ratings (ISLPR) is in their unequal implementation of the three stages. The tests offer measures and reports but may be light on the first stage, description. The scales provide description and reports but may lack a measuring instrument (Davies 1995).

The move away in recent times from the objective test to the subjective scale is no doubt part of the widespread rejection in the social sciences of positivism, fuelled by the sociocultural turn and concern for critical language testing (Shohamy 1997)[1]. But it also has a more practical explanation. In large-scale operations, common standards may be more readily acceptable if they are imposed by a scale which is open to local interpretations. A contemporary example is found in the Council of Europe's Common European Framework of Reference (CEFR) for Languages: Learning, Teaching, Assessment (CEFR 2003). The CEFR, then, is not a measure. For measuring purposes, the CEFR operates as a common reference to which local and national assessment instruments can relate (Taylor 2004).

Large-scale operations like the CEFR may be manipulated unthinkingly by juggernaut-like centralizing institutions. Mitchell describes the misconceived imposition of the attainment targets and level descriptors of the UK's National Curriculum for Modern Foreign Languages, asserting that the longer-term impact of these standards "will certainly be to reduce diversity and experimentation . . . we are likely to lose the more ambitious and more experiential interpretations of communicative language teaching, which has . . . historically been found at local level" (Mitchell 2001, p. 174). Elder reports a similar case of inappropriate standards for Languages Other Than English (LOTE) in Australia (Elder 2000b). Bailey and Butler, discussing the No Child Left Behind (NCLB) program in the USA, complain that, because of recent changes to the federal law, no distinction is made between English learners and native speakers. The law now requires "the inclusion of English Learner

[1]Special Issues of Language Testing (14/3: 1997) and of Language Assessment Quarterly (1/2&3: 2004) are dedicated to the issue of ethics in language testing.

students in new mandated assessment systems. The NCLB Act of 2001 increases school accountability" (Bailey and Butler 2004, p. 183). Such mismatches are not wholly unlike what we have suggested as the possible CEFR massaging of local measures since in all cases, what is in train is the imposition of one overall set of standards nationally, regionally, or even universally, the McDonaldization of language standards. However, our scepticism may be misjudged and out of place, since by their very nature standards are ambitious for wider and wider acceptance. There really is little point, after all, in establishing standards just for me if they have no meaning or application for you or anyone else: similarly with standards for a class, school, city, and so on. What then is wrong about the Mitchell, the Elder, and the Bailey and Butler cases is not that they were attempts at expanding the range and distribution of standards but that they were, for the populations discussed, the wrong standards.

In the second sense, standards are a set of principles which can be used as a basis for evaluating what language testers do, such as carrying out the appropriate procedures. When a school principal maintains that his/her school is "maintaining standards," the implication is that achievement levels over time are constant. When an examination body such as Educational Testing Service (ETS) or the University of Cambridge Local Examinations Syndicate (UCLES) claims that they are "maintaining standards," what they seem to mean is that they are carrying out the appropriate procedures, such as standard setting (Griffin 2001).

Standard setting is a technical exercise, involving, as it does, the determining of cut scores for a test, either for pass/fail or for each level in a band system. But it is worth remembering that standard setting remains a substantially political and ethical issue: "there can be no purely technical solution to the problem of standard setting in this context' (of an English test for ESL health professionals), the decision 'remains intrinsically ethical and political; no amount of technical sophistication will remove the necessity for such decisions." (Lumley et al. 1994, p. 39.)

To an extent, this is where Messick's theorizing (1989) has taken us in his attempt to provide one overall coherent framework for the description, the measurement and the reporting of standards, and the systematic effects they have on all stakeholders. The term that has come to be associated with his conceptualization is, as we have seen, that of consequential validity, but it does seem that impact may be an alternative name for it (Hawkey 2006). Impact studies the effects that a test has when put to use: this is more than the more frequently used term washback precisely because it is concerned not with just how a test works in one situation but with its systemic influences. As such, impact can investigate fundamental issues about standards: are they the right ones for the purposes intended, are they fully and openly described, are they attached to reliable and valid measures and is the reporting clear and precise and does the test produce desirable outcomes in the form of more appropriate and useful teaching? What impact studies, then, can do is to enable us to reevaluate and make explicit not just the standards we promote but the very view of language we take for granted.

## Major Contributions: Ethics

The basic concerns of language testing, that its work should be reliable, valid, and practical and that it should take responsibility for its impact, fall on different sides of the two explanations for acting ethically. The first explanation is the deontological (following the philosopher, Kant) which takes account of its intrinsic value, and the second is teleological (following the philosopher, Hobbes) which takes account only of consequences. Thus we could crudely align validity and practicality with the deontological explanation and reliability and impact with the teleological explanation, which means that ethically we cannot choose between the deontological and the teleological because we must take account of both present value and future effects (Lynch 1997; Kunnan 2000).

One of the chief roles for ethics is to maintain a balance between the rights of the individual and the demands of the social. The danger is that in our attempts to be fair to individuals, we may end up by destroying the social, making all morality individual and therefore never achieving fairness. We are left, writes Osborne, only with "personal ethics or the search for small forms of valid knowledge" (1992, p. 181).

But there is a way out of such a solipsist trap, as Jackson (1996) shows. Discussing Codes of Practice, she points out that morality is never absolute. For example, codes of health and safety require a clause which limits the protection of employees to "within reason." Such a clause takes a common sense approach recognizing that (1) there are rules and (2) how they are interpreted will depend on the local context. For ILTA, this has raised the difficult problem of reconciling its global statement of ethical commitment with what may be differently interpreted in local situations. Hence, the recourse to a twin approach, the Code of Ethics as the statement of abstract principles and the Code of Practice as the explanation of how these principles, is put into local practice.

Codes of Ethics have greater validity for organizations claiming to be professional when there is a single form of activity, one basic qualification, where there is mainly one type of work and where the activity is already strongly organized and formally registered. The professions of law and medicine are again the obvious canonical examples.

It has been suggested that ethics in language testing is no more than an extended validity. This is the argument of Alderson et al. (1995), that ethics is made up of a combination of validity and washback. Validity, and particularly consequential validity, is defined by Messick (1989) as being concerned with the social consequences of test use and how test interpretations are arrived at. Gipps (1994) considers that consequential validity represents a shift from "a purely technical perspective to a test-use perspective – which I would characterize as an ethical perspective" (Gipps 1994, p. 146).

An ethical perspective for a language tester is necessary (Kunnan 2005). But in all professional statements of morality, a limit needs to be imposed on what is

achievable or even perhaps desirable. In my view, therefore, the apparent open-ended offer of consequential validity goes too far. It is just not possible for a tester as a member of a profession to take account of all possible social consequences (Davies 1997a, 2005, 2008). What can be done is the internal (technical) bias analysis and a willingness to be accountable for a test's *fairness* or, in other words, limited and predictable social consequences we can take account of and regard ourselves as responsible for. A language test to select trainees for an organization of torturers is surely unacceptable to the profession. But if that organization, unknown to me, makes use for selection purposes of a language test I have designed, it is surely unjust that I should be deemed guilty or that I should blame myself for unethical conduct. After all, can an architect be blamed if the building she/he has designed some years ago is used for ugly racist meetings?

In the absence of sanctions for exclusion of members for unethical conduct and of the legal backing to require that those who practise language testing are properly qualified and certified, what the professional associations of language testing can offer is to create an *ethical milieu* (Homan 1991) through education: the community of self-governing scholars are inspired deontologically by their ambition to contribute to the public good. Helping create the ethical milieu is the Code of Ethics (and/or of Practice) which makes the direct link between the members of the profession (test developers in all their manifold activities) and their stakeholders. Professionalism is thus demonstrated, and the profession is shown to be accountable by the acceptance of a Code of Ethics, by the publication of the profession's standards and by the recognition of stakeholders' rights.

## Codes of Ethics and of Practice

Codes of Ethics and of Practice across all sectors have proliferated in recent years. This increase raises three questions:

1. Why has there been such a rapid increase?
2. Has the increase improved ethical standards?
3. Do the Codes provide protection for the profession from misuse of their products?

"Ethics codes," write Leach and Oakland (2007), "are designed to protect the public by prescribing behaviors professionals are expected to exhibit." And their spread is clear: "of the two hundred largest corporations in the world, 52.5% have some sort of ethical code" (Helin and Sandstrom 2007, p. 253). However, these same authors conclude at the end of a review of corporate codes of ethics: "we still lack knowledge on how codes work, how they are communicated and how they are transformed inside organizations" (ib, p. 253).

Language testing has in the last 30 years or so sought to professionalize itself. To that end, it has provided itself with both national and international professional associations such as the International Language Testing Association (ILTA CoE 2000), the Association of Language Testers of Europe (ALTE 2001), the European

Association for Language Testing and Assessment (EALTA 2006), and the Japan Language Testing Association (JALT) and three regional associations in the USA, the Midwest Association of Language Testers (MWALT), the East Coast Organization of Language Testers (ECOLT), and the Southern California Association for Language Assessment Research (SCALAR).

A professional Code of Ethics is a set of principles which draws upon moral philosophy and serves to guide good professional conduct. It is neither a statute nor a regulation, and it does not provide guidelines for practice, but it is intended to offer a benchmark of satisfactory ethical behaviour by members of the profession. A Code of Ethics is based on a blend of the principles of beneficence, non-maleficence, justice and respect for autonomy and for civil society.

## ILTA

When ILTA was established in the early 1990s, one of the early projects was to develop a Code of Standards (also known as a Code of Practice). A draft Code was produced in 1997 but the project was not taken further, largely, it seems, because it appeared too difficult to agree on a single ILTA code. ILTA may have been a small organization, but it had a global membership and therefore wished to reach agreement on a single – global – Code. Somewhat later the project was restarted. It was decided that in the first instance a Code of Ethics (CoE) should be developed and not a Code of Practice (CoP) on the grounds that it would be more abstract and therefore more likely to gain universal acceptance. The CoE was developed and accepted by ILTA as its CoE in 2000.

The ILTA Code of Ethics justifies itself thus:

> (it) . . . "is a set of principles which draws upon moral philosophy and serves to guide good moral conduct. It is neither a statute nor a regulation, and it does not provide guidelines for practice, but it is intended to offer a benchmark of satisfactory ethical behaviours by all language testers."

It mentions sanctions and it makes clear that good professional behaviour is dependent on judgment; there are no formal rules and what the Code of Ethics relies on in the absence of sanctions is the willingness of ILTA members to act responsibly in accordance with the Code of Ethics (CoE) because they are professionals. In other words, professional training and experience equip you to behave responsibly. Those who fall short may be stripped of their ILTA membership. That mirrors the procedure in law and medicine, but in those professions the sanctions are very much more effective. Without membership of the relevant legal and medical professional bodies, it is not possible to practise as a lawyer or a doctor. That is just not the case in language testing where the sanctions are weak and not supported by the law. Thus, there is nothing to prevent an ex member of ILTA to continue to practise as a language tester. While the law and medicine are strong professions, language testing

is a weak profession where the burden of being professional is more an individual than a collective responsibility.

The ILTA Code of Ethics identifies nine fundamental principles, each elaborated on by a series of annotations which generally clarify the nature of the principles; they prescribe what ILTA members ought to do or not do, or more generally how they ought to comport themselves or what they, or the profession, ought to aspire to; and they identify the difficulties and exceptions inherent in the application of the principles. The annotations further elaborate the Code's sanctions, making clear that failure to uphold the Code may have serious penalties, such as withdrawal of ILTA membership on the advice of the ILTA Ethics Committee. Although this Code derives from other similar ethical codes (stretching back into history), it does endeavor to reflect the ever-changing balance of societal and cultural values across the world.

Language testers are independent moral agents, and sometimes they may have a personal moral stance which conflicts with participation in certain procedures. They are morally entitled to refuse to participate in procedures which would violate personal moral belief. Language testers accepting employment positions where they foresee may be called on to be involved in situations at variance with their beliefs have a responsibility to acquaint their employer or prospective employer with this fact. Employers and colleagues have a responsibility to ensure that such language testers are not discriminated against in their workplace.

## Three Questions

I return now to the three questions I posed earlier. The first is: why has there been such a rapid increase (in the publication of Codes of Ethics)?

It has been suggested that everyone today wants to be a professional, that every work activity now desires to professionalize itself for purposes of prestige and to secure greater control over those involved in the activity. Indeed, in Western societies it could be argued that the familiar distinction between professions and trade unions is now blurred. And as professions have multiplied, so have Codes of Ethics or Practice by the newer professions, anxious to claim their status as a profession and to do so in the public way that publishing a Code permits.

Siggins comments: "Codes of ethics and codes of practice have multiplied without pause in the last decades of this century, not only in the professions but in business, industry and social services, largely in response to the successful growth of consumer movements and their demand for accountability to the public interest" (Siggins 1996, p. 53). He goes on to explain that those commercial activities that have issued codes are wholly concerned "to acknowledge legal and moral rights of their customers and their duty towards them. . . Codes of the learned professions, on the other hand, have always declared the virtue and competence of the select members of a distinguished class" (ib.).

The code takes its origin in the oath taken in law and medicine. "When in the nineteenth century, medical associations in Britain, the US, Australia and elsewhere called their corporate ethical standards 'code of ethics', they intended to echo the

prescriptive force of this usage to express what the Hippocratic oath had already called the 'law of medicine'" (: ib., p. 156). "The learned professions' use of codes always differed from the more recent commercial use by their declaration of the virtue and competence of their members who are 'select members of a distinguished class'" (p. 55). For language testing, caught up in their rush to professionalize, the question has been: Is language testing a *learned* profession or a social/commercial activity? The answer has to be that it is both: for those members who are academics, they see themselves as belonging to a learned profession, while those in more business-like bodies, they regard themselves (or perhaps their organizations) as more commercial.

I turn now to Question 2: Has the increase in Codes improved ethical standards?

There is no easy answer to this question. What can be said is that: "A professional grouping risks being characterised as unethical if it does not now espouse a set of principles enshrined in a Code of Ethics" (Freckelton 1996, p. 131). What a code does is to clarify to the members of the profession what it stands for – it acts as a unifying statement; at the same time, it makes clear to the public what may be expected of members of this profession. It is, indeed, a modern version of an oath. However, codes have their critics, that they are elitist and exclusive, that they are hypocritical by claiming what no one in reality practises that they act as good public relations, and that they provide the profession with a moral screen to hide behind; once the code has been published, it can be set aside and ignored, while "professional" practice continues as brokenly as before (Davies 2004; Boyd and Davies 2002).

It is necessary to emphasize that a Code of Ethics or Practice, or Guidelines for Practice or Ethical Standards are not rules, and they are certainly not laws. The most we can expect of them is that they "make a contribution to improving behaviour in the areas they deal with" (Coady 1996, p. 287). Where there are sanctions leading to loss of membership and subsequent inability to practise the profession (as in law and medicine), then the Code comes nearer to a law. But this is not possible for most professions, and it is not possible, so far, for language testing.

Question 3: Do the Codes provide protection for the profession from misuse of their products?

Tests are not developed or used in a value-free psychometric test tube; they are virtually always intended to serve the needs of an educational system or of society at large. Cronbach (1984) has succinctly stated this fact as follows: "testing of abilities has always been intended as an impartial way to perform a political function – that of determining who gets what" (Cronbach 1984, p. 5; Bachman 1990, pp. 279, 280). And Spolsky has no doubt about the main purpose of tests, stating that "[f]rom the beginning, public tests and examinations were instruments of policy" (Spolsky 2009, p. vii). Tests, then, are used for political purposes, language tests as much as, perhaps more than, other tests, they perform a social function. Those who use tests are being political: the question I address in this section is who makes those political decisions about the use of language tests, and to what extent those language testers who develop tests are responsible, ethically if not legally, for that use (Shohamy 2001).

Fulcher and Davidson (2007) agree with Messick (1989) that decisions about the test use need to be considered in terms of their consequences, very much a teleological approach: "The intention of a decision should be to maximize the good for the democratic society in which we live, and all the individuals within it. . . . we may define any test as its consequences" (Fulcher and Davidson 2007, pp. 142, 143). In the case of the infamous Dictation test (Davies 2004; McNamara 2006), employed in Australia in the first half of the twentieth century for the purpose of excluding unwanted immigrants, those who developed the test like those who used it were clearly responsible, ethically responsible, because its explicit use as a test was intended. As Fulcher and Davidson (2007) write:

> . . . an unintended use may take one of two forms: (1) an unintended use that the test developers do not know about or do not approve of, or (2) an unintended use that the test developers know about and approve of. . . . Both are equally invalid unless a new argument is constructed for the new testing purpose and evidence collected to show that the retrofitting is valid, so that the same test may be shown to be useful in a new decision-making context . . . retrofitting test purpose without the construction and investigation of a new validity and utilization argument constitutes an implicit claim that any test can be used for any purpose, which is to introduce validity chaos. (ib., p. 175)

McNamara and Roever (2006) consider a range of language tests used for establishing social identity, ranging from the celebrated biblical story of the shibboleth (Lado 1949; McNamara 2005; Spolsky 1995) to present-day language tests used to determine the claims of asylum seekers. They write: "the politics and ethics of the use of these tests are complex. . . The procedures involved are widely used in Europe, Australia and elsewhere in the pioneering of claims of undocumented asylum seekers . . . the lack of validity considerations in their implementation leads to serious injustice, a situation that would be remedied in large part by attention to the quality of the testing procedure" (McNamara and Roever 2006, p. 165). The objection, then, that McNamara and Roever have to these procedures, which involve assessing whether the claimant really does come from the country which she/he claims to come from by matching his/her accent/dialect to that country or region, is in terms of their validity. They criticize the sociolinguistic construct which all too commonly assumes a homogeneity of accent/dialect where none exists, and they criticize the methods used in the assessment. In other words, they have no objection to the testing of asylum seekers to determine their honesty. This is what they write: "Although some applied linguists and language testers have objected to the use of such procedures altogether, it is reasonable to think that the evidence that they provide, when that evidence is properly obtained and interpretable, might be as useful in supporting a valid claim to asylum as in denying an invalid one" (p. 172).

McNamara and Roever do not object. They refer to the guidelines for the proper use of language analysis in relations to questions of national origin in refugee status (Language and National Origin Group 2004). Where does this leave their assertion, already quoted, that "the politics and ethics of these tests are complex" (p. 165)? The politics of their argument is straightforward: it concerns the national decision to offer asylum only to those who are genuine refugees and to exclude those who aren't. The

implementation of that intention in the procedures they discuss is a matter of validity, and, as they show, they fail that test. Where, then, are the ethical concerns? Presumably they have to do with the use of such procedures and are a judgment on the exclusion measures. As McNamara and Roever have shown, there is no agreement on this in the language testing profession – and it may be (as the ILTA Code of Ethics makes clear) that while the profession accepts the need for such testing, which is, after all, quite legal, there will be acceptance of those individuals in the profession who choose not to participate on grounds of conscience. So are these tests ethical? It would seem that, according to McNamara and Roever, they are potentially externally valid, but they lack internal ethicality, that is, they are not valid.

Can a language test be fair and is fairness an ethical consideration? McNamara and Roever (2006) discuss fairness in the context of the ETS Fairness Review Guidelines (ETS 2003) and of the various Codes of Ethics and Practice referred to earlier in this article. They recognize the difficulty of setting a global norm for fairness (p. 137). Fairness, McNamara and Roever propose, is a professional obligation. If fairness is an ethical component, they are right. But what exactly is fairness? Examining Rawls (2001) on fairness and justice, Davies (2010) argues that in language testing it is validity rather than fairness which must be the criterion:

> A test that is valid for group A (say adults) cannot be valid for group B (say children) because they belong to different populations. It is not whether such a test is fair or unfair for group B: the test is just invalid for group B. The search for test fairness is chimerical." (Davies 2010, p. 175)

This leads back to the earlier discussion on language testing for asylum seekers and raises the issue of language tests for citizenship. The proposed UK legislation for pre-entry language tests was debated in the House of Lords on 25 October 2010. Briefing points were quoted from Adrian Blackledge of the University of Birmingham who argued that such tests were not valid for purpose. Charles Alderson was also mentioned in the debate. He commented that "the UK Border Agency's August 2010 list of approved providers of the English test has been developed by unknown agencies with absolutely no evidence of their validity, reliability etc." (Hansard 25 Oct 2010, pp. 1101, 1102) (ILPA 2010).

These two critics approach the issue from the two different ethical positions discussed earlier, one from the point of view of ethics for use (no test for this purpose could be valid) and the other from the point of view of the internal validity of the test (this test lacks the necessary requirements of a language test). What Alderson appears to be claiming here – unlike Blackledge, but like McNamara on the testing of asylum seekers – is that such a test could be ethical if it were a satisfactory test.

The moral philosopher, Peter Singer, contends: "what is it to make moral judgments or to argue about an ethical issue or to live according to ethical standards? Why do we regard a woman's decision to have an abortion as raising an ethical issue but not her decision to change her job?" (Singer 2002, p. 13). Singer's answer is the golden rule: an ethical belief or action or decision is one that is based on a belief that

it is right to do what is being done. Ethics, Singer argues, is a set of social practices that has a purpose, namely the promotion of the common welfare. "Moral reasoning, therefore, is simply a matter of trying to find out what is best for everyone, achieving the good of everyone alike – the golden mean" (Davies 2004, p. 98).

Professional ethics, therefore, is about the ethics of the profession, not about morality which is a matter for the individual. In becoming a member of a profession, the new entrant agrees to uphold the ethics of the group – with which his/her own conscience may not always agree. The various Codes (of Ethics, of Practice, of Standards. . .) make public what it is members are prepared to agree to, what it is they *swear* by, and they reach this agreement through compromise. They accept responsibility for the development of the language tests they work on and for the intended consequences of those tests. But they do not accept responsibility for any unintended consequences, nor should they.

The prisoner's dilemma in game theory presents the role, both theoretical and practical, of ethics, placing the emphasis on the importance of being unselfish: two prisoners, A and B, are arrested for complicity in the commission of a crime (they are in fact both guilty). They are put in cells between which no communication is possible and then offered a deal. The deal is as follows:

1. If A confesses and B does not (or vice versa), then A is released and B gets 10 years.
2. If both A and B confess, they each get 5 years in prison.
3. If neither confesses, they each get 1 year in prison.

The best (selfish) strategy for A is to confess. Then, if B does not confess, B gets 10 years and A is released. However, A does not know what B will do: it is possible that B will also confess, in which case, they both get 5 years. The best (selfish) strategy might therefore not work, indeed it could work to A's disadvantage. The best result would be obtained if neither A nor B confesses. However, this is still risky as a strategy for A since B may confess, in which case A would get 10 years and B be released. What is necessary is for both A and B to think not of the best strategy for themselves alone (the selfish approach) but of the best outcome for them both (for *the profession*). If they each take concern for the other then neither will confess, in which case they will both get 1 year (Davies 1997b, pp. 329, 330).

Discussing this dilemma, Scriven (1991) concludes:

> The only solution is through prior irreversible commitment to treat the welfare of each other as comparable to their own, and this reflects the reason for ethical training of the young'. Being ethical comes at a cost to oneself (both A and B would have to go to prison for 1 year) but for the group/society/company/profession etc., the cost is worth while since an ethical society has better survival rate value than a society of rational egoists. (Scriven 1991, p. 276)

And yet, the ethical dilemma remains. What the ethical imperative for the profession does is to ensure the best results for the profession. But professions, as we know, do not always behave ethically. The law, the legal profession, after all, for

long maintained the right of some members of society to own slaves who were regarded not as citizens but as property. Singer's appeal to the golden mean as the best for everyone, achieving the good of everyone alike, is surely desirable only if "everyone" means everyone.

## Problems and Difficulties: Rights

Rights are of two kinds, natural or inalienable rights and civil rights. Natural rights are those freedoms which belong to every individual by virtue of being human: they include the right to protect one's life and property. Civil rights include those rights granted to the citizens of a state by its legal institutions and legislative authorities. These imply the right of access to the legal system for protection and claims against others, for defense against charges, for protection of the law, and for equality of treatment under the law.

There is no absolute distinction between natural and civil rights. Claims (or needs or aspirations) such as a good education, decent housing, health care, employment, an adequate standard of living, equality of opportunity, freedom of speech, and freedom to take part in political processes are regarded by some as belonging to civil rights and by others to natural rights. Furthermore, since it is governments that protect and maintain (and, pragmatically, grant) all rights, then even inalienable rights may be (and sometimes are) regarded as kinds of civil rights. The argument is that a right which has not been granted by the state is not a real right; thus, in a slave-owning society, slaves, it could be argued, have no natural right to freedom or equality because their society does not accord equal rights as citizens to slaves. However, the point of making a distinction between civil and natural rights is to make explicit that some rights (such as political participation, equality under the law) remain rights, they are inalienable even if they are not granted, even if they are removed.

Rights do not exist on their own. They impose reciprocal obligations, duties to act in certain ways as required by moral or ethical principles, promises, social commitments, and the law. My claim to a right requires that I accept that it imposes an obligation. For example, my right to free speech means that I acknowledge that others also have the same right and that I accept that in pursuing my right I do not harm others' rights to, for example, the pursuit of their own happiness or their right to equal treatment under the law. In other words, I must not in exercising my right to free speech tell lies about other people. Further, I must accept that my right to free speech, for example, entails an acceptance on my part that in order to fulfil my obligation to others, I must be prepared to limit my own right.

The universality of human rights requires that everyone act to ensure that others' rights are also observed. However, the new professionalism, influenced no doubt by the climate of postmodernism, is about giving more power to users in the context of the professional relationship, even though the focus is still on the professional as the one giving the power (Banks 1995, p. 105). In the same way, the critical turn in

applied linguistics and language testing (Pennycook 2001; Shohamy 2001) insists on the ethical importance of recognizing the rights of all stakeholders. And since language testing disempowers test takers, the ILTA Code of Practice highlights test takers' rights.

## Conclusions and Future Directions

Being professional, a state to which many aspire, means making a commitment to ethics, establishing and observing standards and recognizing the rights of all those professionals engage with, including themselves. A profession becomes strong and ethical precisely by being professional (Davies 1997b). What a Code of Ethics does is to remind us of what we already know, that language testers are a serious organization, committed to a social purpose, to maintaining standards, to upholding the rights of all stakeholders, and to working professionally with colleagues. It is important to spell out in a Code of Ethics what this means, but there is something to be said for the conclusion that Alderson, Clapham, and Wall (1995) came to, that being ethical in language testing could be guaranteed by the traditional precepts of reliability and validity.

As for responsibility for test use, this must be limited, as Fulcher and Davidson (2007) point out, to the purpose for which the designer has validated the test. Where does this leave tests for asylum and citizenship, and the use by government agencies of invalid tests (ILPA 2010)? In the case of the first (asylum and citizenship), what is ethical in language testing – what the Codes require – is that the tests should be properly designed, valid for their purpose. The profession does not oppose such tests. However, as the ILTA Code of Ethics makes clear, while not opposing such tests, the profession does not require members to take part in their construction if they have a conscientious objection against them. The imposition of such tests is a political matter, and the Codes have nothing to say about politics. What is ethical for the profession is not necessarily moral for every individual member. In the case of the second (government use of invalid tests), the Codes again insist that it is professionally irresponsible to use invalid tests. However, correct though that argument is, it can succeed only if government and other agencies are willing to heed professional advice. Otherwise, like the Australian government's attitude to the Dictation Test, what decides is politics and not ethics.

## Cross-References

► Critical Language Testing
► History of Language Testing
► Washback, Impact, and Consequences Revisited

## Related Articles in the Encyclopedia of Language and Education

Stephen May: Language Education, Pluralism, and Citizenship. In Volume: Language Policy and Political Issues in Education

Bonny Norton, Ron David: Identity, Language Learning and Critical Pedagogies in Digital Times. In Volume: Language Awareness and Multilingualism

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

## References


Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

ALTE. (2001). The association of language testers of Europe code of practice. http://www.alte.org

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bailey, A. L., & Butler, F. A. (2004). Ethical considerations in the assessment of the language and content knowledge of English language learners K-12. *Language Assessment Quarterly, 1*(2 & 3), 177–193.

Banks, S. (1995). *Ethics and values in social work*. London: Macmillan Press.

Boyd, K., & Davies, A. (2002). Doctors' orders for language testers: The origin and purpose of ethical codes. *Language Testing, 19*, 296–322.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing, 15*(1), 45–85.

CEFR. (2003). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR). Manual: Pilot preliminary version*. Council of Europe, Strasbourg: Language Policy Division.

Coady, C. A. J. (1996). On regulating ethics. In M. Coady & S. Bloch (Eds.), *Codes of ethics and the professions* (pp. 269–287). Melbourne: Melbourne University Press.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper and Row.

Davies, A. (1995). Introduction: Measures and reports. *University of Melbourne: Melbourne Papers in Language Testing, 4*(2), 1–11.

Davies, A. (1997a). Introduction: The limits of ethics in language testing. *Language Testing, 14*(3), 235–241.

Davies, A. (1997b). Demands of being professional in language testing. *Language Testing, 14*(3), 328–339.

Davies, A. (2004). Introduction: Language testing and the golden rule. *Language Assessment Quarterly, 1*(2&3), 97–107.

Davies, A. (2005). *A glossary of applied linguistics*. Edinburgh/New Jersey: Edinburgh University Press and Mahwah/Lawrence Erlbaum Associates.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education (2nd ed.) volume 7: Language testing and assessment* (pp. 429–443). New York: Springer.

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2), 171–176.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: University of Cambridge Local Examinations Syndicate and University of Cambridge.

EALTA. (2006). European Association for Language Testing and Assessment. *EALTA Guidelines for good practice in language testing and assessment*. http://www.ealta.eu.org. Accessed 30 Oct 2010.

Elder, C. (2000a). Preface. In C. Elder (Ed.), Defining standards and monitoring progress in languages other than English. Guest edited issue of the *Australian Review of Applied Linguistics, 23*(2), 1–5.

Elder, C. (2000b). Learner diversity and its implications for outcomes-based assessment. In Elder, C. (Ed.), Defining standards and monitoring progress in languages other than English. Guest edited issue of the *Australian Review of Applied Linguistics, 23*(2), 36–61.

ETS (Educational Testing Service). (2003). Fairness review guidelines. Princeton: author: available from http://www.ets.org/Media/About ETS/pdf/overview.pdf

Freckelton, I. (1996). Enforcement of ethics. In M. Coady & S. Bloch (Eds.), *Codes of ethics and the professions* (pp. 130–165). Melbourne: Melbourne University Press.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Fullwinder, R. K. (1996). Professional codes and moral understanding. In M. Coady & S. Bloch (Eds.), *Codes of ethics and the professions* (pp. 72–87). Melbourne: Melbourne University Press.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Griffin, P. (2001). Establishing meaningful language test scores for selection and placement. In C. Elder, A. Brown, K. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in Honour of Alan Davies* (pp. 97–107). Cambridge: Cambridge University Press.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Helin, S., & Sandstrom, J. (2007). An inquiry into the study of corporate codes of ethics. *Journal of Business Ethics, 75*, 253–271.

Homan, R. (1991). *The ethics of social research*. London: Longman.

Illich, I. (1987). *Disabling professions*. New York: M. Boyars.

ILPA: Immigration Law Practitioners' Association. (2010). House of Lords Motion re: Statement of Changes in Immigration Rules (Cm 7944) 25 Oct 2010. www.ilpa.org.uk. Accessed 30 Oct 2010.

ILTA CoE. (2000). International Language Testing Association Code of Ethics. http://www.iltaonline.com. Accessed 30 Oct 2010.

Jackson, J. (1996). *An introduction to business ethics*. Oxford: Blackwell.

Kunnan, A. J. (2000). *Fairness and validation in language assessment*. Cambridge: UCLES/ Cambridge University Press.

Kunnan, A. J. (2005). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 779–794). Mahwah: Lawrence Erlbaum.

Lado, R. (1949). Measurement in english as a foreign language, Unpublished PhD dissertation, University of Michigan

Language and National Origin Group. (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *The International Journal of Speech, Language and the Law, 11*(2), 261–166.

Leach, M. M., & Oakland, T. (2007). Ethical standards impacting test development and use: A review of 31 ethical codes impacting practices in 35 countries. *International Journal of Testing, 7*(1), 71–88.

Lumley, T., Lynch, B., & McNamara, T. (1994). Are raters' judgements of language teacher effectiveness wholly language based? *University of Melbourne: Melbourne Papers in Language Testing, 3*(2), 40–59.

Lynch, B. K. (1997). In search of the ethical test. *Language Testing, 14*(3), 328–339.

Marshall, G. (Ed.). (1994). *The concise Oxford dictionaty of sociology*. Oxford: Oxford University Press.

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 1–20.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31–51.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, and Macmillan.

Mitchell, R. (2001). Prescribed language standards and foreign language classroom practice: Relationships and consequences. In C. Elder, A. Brown, K. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in Honour of Alan Davies* (pp. 163–176). Cambridge: Cambridge University Press.

Osborne, R. (1992). *Philosophy for beginners*. New York: Writers and Readers Publishing.

Pennycook, A. (2001). *Critical applied linguistics: A critical introduction*. Mahwah: Lawrence Erlbaum.

Rawls, J. (2001). *Justice as fairness: A restatement*. Cambridge, MA: Harvard University Press.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park: Sage.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*, 340–349.

Shohamy, E. (2001). *The power of tests: A critical perspective on the use of language tests*. Harlow: Longman.

Siggins, I. (1996). Professional codes: Some historical antecedents'. In M. Coady & S. Bloch (Eds.), *Codes of ethics and the professions* (pp. 55–71). Melbourne: Melbourne University Press.

Singer, P. (2002). *Writings on an ethical life*. New York: Harper Collins.

Skene, L. (1996). A legal perspective on codes of ethics. In M. Coady & S. Bloch (Eds.), *Codes of ethics and the professions* (pp. 111–129). Melbourne: Melbourne University Press.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Spolsky, B. (2009). Editor's introduction. In *Annual review of applied linguistics* (Vol. 29, pp. vii–xii). Language Policy and Language Assessment.

Taylor, L. (2004). *Introduction: Issues of test comparability* (Research notes, Vol. 15). Cambridge: UCLES.

Weber, M. (1948). From. In C.Wright Mills and H. H.Gerth (Ed. & Trans.), *Max weber: Essays in sociology*. London: Routledge and Kegan Paul.

# The Common European Framework of Reference (CEFR)

Monica Barni and Luisa Salvati

**Abstract**

Drawn up by the Council of Europe, published in 2001 and adopted by the European Commission in its language policies, the CEFR was born within a scenario where the EU member states and its institutions were characterized by linguistic and cultural diversity. Since 2001 the CEFR has become the most important reference document in the fields of language learning, teaching, and assessment, both in Europe and beyond, but in recent years, little attention has been paid to the debate concerning the direct impact of the CEFR on language teaching and assessment and, consequently, on language policy throughout the EU.

Indeed, although the theoretical approach of the CEFR reflects the will of the EU to address multilingualism as an asset allowing for the active inclusion of all citizens, CEFR descriptors define the linguistic competence from a monolingual perspective, using arbitrary standards relying on professional experience rather than on empirical data based on actual learner performance. Furthermore, the CEFR and its standards are often used as benchmarks in migrant competence, although they were not created for this, which changed the CEFR from a tool used to measure language knowledge to a political instrument. The Italian situation can be considered emblematic as concerns the lack of reflection on the (mis-)use of the CEFR and the fact that it is too often used as a label without considering the impact and consequences of such a use, according to which CEFR levels are now enshrined in laws and policies incorporating the administration of language tests in migration domains.

M. Barni (✉) • L. Salvati
University for Foreigners of Siena, Siena, Italy
e-mail: barni@unistrasi.it; salvati@unistrasi.it

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_29

**Keywords**
Assessment • CEFR • Plurilingualism

## Contents



## The CEFR: A Historical Overview

The *Common European Framework of Reference for Languages*: *Learning*, *Teaching and Assessment* (henceforth CEFR) was drawn up in 1996 by the Council of Europe which made the document accessible to everybody on its website (Council of Europe 1996). The open-access system on the Council of Europe website had a strong symbolic value, implying the possibility of spreading the language and culture policy document. Published in 2001 both in English and in French editions, the CEFR was conceived with an aim to value and increase the linguistic and cultural diversity the EU member states and its institutions are characterized by, and which, nevertheless, generates a widespread concern about social cohesion and integration.

Based on three main concepts – use, knowledge, and ability – the nine chapters of the CEFR illustrate the theoretical model for describing linguistic competence, the six levels of proficiency, the contexts of language use, the learning process, as well as the operational implications on assessment. The document addresses both those who learn a language and those who are involved in language teaching and assessment.

Notwithstanding this, although the theoretical approach of the CEFR – like most of the EU documents and recommendations (Council of the European Union 2002) – reflects the will of the EU to address multilingualism as an asset allowing for the active inclusion of all citizens, multilingualism is dealt with most often as a problem (Blommaert et al. 2012). This becomes visible in the case of migrants, and not all kinds of multilingualism are considered as having the same value, since no consideration is given to immigrant languages (Extra and Yağmur 2012).

Since 2001, the CEFR has been adopted by the European Commission in its language policies (e.g., Committee of Ministers 2008; EC Action Plan 2004–2006). Its greatest merit lies in providing a valuable guidance for goals, methods, development of curricula, and teaching materials selection, representing not only a theoretical reference system both in Europe and beyond but also a guide to implementation in the field of learning, teaching, and assessment of languages (see North 2001; Morrow 2004; Trim 2010). Nevertheless, no deep reflection has been done in these years about the direct implications the CEFR ideology of language has had on language teaching and assessment and, consequently, on language policy in the

EU. In the next paragraphs, we will try to shed light, with the support of various documents (Council of Europe 2009, among others), on some contradictions of the CEFR compared to its initial intents.

## The Contradictions in the CEFR: Plurilingualism and Monolingualism, Norm and Performance, and Language Testing

The first point on which we intend to develop some reflection concerns the paradox of the plurilingual approach of the CEFR, in contrast with the monolingualism toward which its descriptors tend. Indeed, the CEFR is based on a plurilingual approach, which reflects a pragmatic and sociolinguistic view of language, mainly conceived in a social and interactional dimension:

> The approach adopted here, generally speaking, is an action-oriented one in so far as it views users and learners of a language primarily as 'social agents', i.e. members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action. (Council of Europe 2001, p. 9)

Such a view of language is founded on the idea of communicative language competence, consisting of three components: a linguistic, a sociolinguistic, and a pragmatic one. Linguistic competences "include lexical, phonological, syntactical knowledge and skills and other dimensions of language as system, independently of the sociolinguistic value of its variations and the pragmatic functions of its realizations" (Council of Europe 2001, p. 4). Sociolinguistic competences "refer to the sociocultural conditions of language use" (ibidem): in this component the plurilingual approach finds its fullest expression as it "affects all language communication between representatives of different cultures, even though participants may often be unaware of its influence" (ibidem). Finally, pragmatic competences "are concerned with the functional use of linguistic resources (production of language functions, speech acts), drawing on scenarios or scripts of interactional exchanges" (ibidem).

Therefore, the plurilingual approach of the framework refuses a merely structural concept of competence and embraces the idea of language as a social action. Moreover, the theoretical approach also takes into account the results of research studies on second language acquisition, establishing a relationship between its didactic proposal and the studies on processes of competence development:

> since it is one of the principal functions of the Framework to encourage and enable all the different partners to the language teaching and learning processes to inform others as transparently as possible not only of their aims and objectives but also of the methods they use and the results actually achieved, it seems clear that the Framework cannot confine itself to the knowledge, skills and attitudes learners will need to develop in order to act as competent language users, but must also deal with the processes of language acquisition and learning, as well as with the teaching methodology. (Council of Europe 2001, p. 18)

Nevertheless, the impression one gets from reading some pages of the CEFR is that the idea of a plurilingual approach is only a theory, an empty model, as the operational implications arising from the issues of languages in contact are never clearly defined:

> the plurilingual approach emphasises the fact that as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other people (whether learnt at school or college, or by direct experience), he or she does not keep these languages and cultures in strictly separated mental compartments, but rather builds up a communicative competence to which all knowledge and experience of language contributes and in which languages interrelate and interact. (Council of Europe 2001, p. 4)
>
> move away from the supposed balanced dichotomy established by the customary L1/L2 pairing where by stressing plurilingualism where bilingualism is just one particular case (*ibidem*).
>
> consider that a given individual does not have a collection of distinct and separate competences to communicate depending on the languages he/she knows, but rather a plurilingual and pluricultural encompassing the full range of the languages available to him/her. (Council of Europe 2001, p. 168)

Although the intents of the CEFR, on a theoretical level, seem to mirror the awareness that human interactions bear "the traces of worldwide migration flows and their specific demographic, social and cultural dynamics" (Blommaert and Rampton 2011, p. 2) in a context of "super-diversity," as it has been defined (Vertovec 2006, 2007), or "hyper-diversity" (Baynham and Moyer 2012; Kelly 2008), the linguistic competence is still mainly described from a monolingual perspective. The language descriptors are largely based on a monolingual view according to which only standard language is supposed to be used. Language continues to be considered as a bounded system linked with bounded communities, and a plurilingual repertoire is just considered as the juxtaposition of different monolingualisms (language 1 plus language 2 . . . language *N*); people are still considered plurilingual when they are able to speak different languages and, in interactions, are able to switch from one language to another. Consequently, people with very rich linguistic repertoires and with a learning background far from formal education – like migrants for whom, in most cases, the official language of the host country is an L2, while one, two, or more language (s) and also mixtures of languages are used in their linguistic exchanges – are considered as lacking sufficient competence in "the" language of "the" country, as we will see below.

In this perspective, the CEFR, as Byram and Parmenter (2012) note, has become an operational tool used to justify choices in language policies, both at an educational and a social level:

> the CEFR is clearly a policy document bearing values and intentions. Yet, like any text, the intentions of its authors may not be read by its users and be taken in entirety but only used in part for the purposes of the users. (p. 4)

Moreover, the part of the document that has been read and used the most and that has had the strongest impact on education and society at large is the one describing the scales of proficiency. As a result, the CEFR is mainly seen in terms of levels of proficiency in a language.

However, which language proficiency is considered? This is the second point to which we would pay specific attention. As McNamara (2009) notes, the genesis of frameworks like the CEFR is characterized by several features, among which is the fact that the acceptability of the framework was negotiated by all interested parties during the course of its development and is not based on empirical evidence, as the same is now repeating for the validation of the CEFR descriptors of mediation:

> [. . .] the CEFR is primarily a policy coordination and administrative initiative, acting as an accounting system and management tool whereby control is exercised by specifying the outcomes of learning independently of any specific test (or language; McNamara 2011). Policymakers need tools that serve their need, which are for accountability, administrative ease, ease of explanation to stakeholders, "scientific" respectability, and so on. The CEFR, with its pyramidal shape (culminating in the six numbered reference levels), is such a tool. The functionality of a universal letter/number system to code the six levels is the key feature of the CEFR, which makes it attractive to administrators and policy makers. (p. 227)

In line with McNamara, Chapelle (2012) defined alignments to frameworks such as the CEFR as "controversial because they attempt to connect social and political meanings associated with frameworks with the scientific procedures used to understand score meaning" (p. 25). Therefore, the CEFR would describe the linguistic competence not only from a monolingual point of view but also using standards (Spolsky 2008), relying on professional experience rather than on empirical data based on actual learner performance, in order to produce uniformity despite "the complexity of languages and human behavior" (Cumming 2009).

Indeed, as the CEFR states:

> However, it is not usually advisable to include descriptors of communicative activities in the criteria for an assessor to rate performance in a particular speaking or writing test if one is interested in reporting results in terms of a level of proficiency attained. This is because to report on proficiency, the assessment should not be primarily concerned with any one particular performance, but should rather seek to judge the generalisable competences evidenced by that performance. There may of course be sound educational reasons for focusing on success at completing a given activity, especially with younger Basic Users (Levels A1; A2). Such results will be less generalisable, but generalisability of results is not usually the focus of attention in the earlier stages of language learning. This reinforces the fact that assessments can have many different functions. What is appropriate for one assessment purpose may be inappropriate for another. (Council of Europe 2001, p.168)

With regard to this explanatory note, Harsh (2014) concludes that:

> Although the CEFR scales have been empirically calibrated using teacher judgment (North 2002), this does not amount to a validation of the scales for specific purposes, such as

> assessor-oriented or constructor-oriented purposes. Similarly, the statements in CEFR Chapter 9 (Council of Europe 2001, pp. 180) that proficiency scales might be useful for such purposes would need to be backed up by empirical validation. (p. 161)

The third point of reflection is the political use of the CEFR and, more specifically, the link between the CEFR and its standards of language testing. As an example, the language descriptors of the CEFR are used as benchmarks in migrant competence, although they were not created for them (McNamara 2011). There has been a "shift in the understanding of the functions, status and roles of language tests. From tools used to measure language knowledge, they are viewed today more and more as instruments connected to and embedded in political, social and educational contexts" (Shohamy 2007, p. 117). Language tests are more gradually being used as policy tools for declared and undeclared policies (Shohamy 2006): "establishing entrance criteria that include a test of another language, a new de facto policy is created, the implication of which is that the 'tested' language becomes the most important language to acquire and master" (Shohamy 2007, p. 120). In addition to this, criteria and constructs of tests embody and sustain the most appropriate language variety that should be used by people ("the" norm), imposing, in this sense, monolingual policies and consequently suppressing multilingual diversity.

As Shohamy states (2004), the implications of the political use of language tests involve determining a hierarchy of languages, suppressing diversity, homogenizing languages, and perpetuating criteria of correctness:

> One of the most salient uses of tests affecting language policies is perpetuating language homogeneity, a construct which is detached from the reality of what languages are and how they are being used, especially in multilingual societies. Most tests impose homogeneous criteria of correctness and thus deliver a message that languages are uniform, standard and follow the same written norms. (Shohamy 2007, p. 124)

Therefore, the power of tests becomes even stronger when test criteria affect language policy, and the definitions of "what it means to know a language" answer generic descriptions which are far from any context and from the contextualized nature of language and language performance in multilingual scenarios.

As a case of how tests become more evident and problematic when they are used as gatekeepers, we will describe the Italian situation. An example of the political use of the CEFR and its levels is the Ministerial Decree of June 4, 2010, which introduces a test in Italian (the level chosen for immigrants is A2) for those migrants who request a long-term residence permit. Among the reasons behind the introduction of the test, the CEFR itself is cited in the preamble to the decree, as a document that is believed to give the mandate for such tests. According to the same agreement, the tests are to be implemented, administered, and assessed by teachers in each of the adult education accredited centers in Italy. This choice implies that dozens of different tests and markers are used, and everything is justified in the name of the CEFR and European language policies, but all decisions are left to individual teachers.

In support of what we state, between December 2010 and May 2011 (the most recent available data), the average rate of failure on the test was 13.6% nationally. The city of Turin showed the highest number of candidates passing (96.5% pass rate) and similar figures were seen in Rome (96%) and Naples (95%). Much lower passing rates were seen in Milan (86%), Venice (70%), and Verona, where only 65% passed the test (Masillo 2015).

At this point, it is reasonable to wonder whether these results reflect immigrants' proficiency in Italian or they are influenced by the characteristics of the person conducting the assessment, where such characteristics involve not only theoretical knowledge and technical skills about language testing and assessment but also the attitude toward the persons being passed. Such a situation can be emblematic of the consequences of such a use of the CEFR, as Van Avermaet (2008) underlines, dealing with language and societal knowledge regarded as a key element in integration, education, and language policies:

> In a policy of a more conditional nature, language courses and language tests have to be more uniform in format as well as in content. A universal and fixed level of language proficiency for all immigrants is a prerequisite. In an obligatory policy, failure or success in a language course or language test can function as a gatekeeper, a mechanism to exclude people.
>
> In a more facilitating policy, language courses and language tests can be more flexible, more tailor made in format and content. The level of language proficiency can vary depending on the needs of the immigrants and on the linguistic requirement in specific domains of the host society in which an immigrant wants to function. A more facilitating policy is more encouraging than discouraging. It is aimed at integration and non-discrimination. It also offers more opportunities for acknowledgement of immigrants' plurilingual repertoires.

These and other questions, therefore, continue to foster the need for reflection on the CEFR and its impact on language assessment and language policies.

## Conclusion: Problems, Difficulties, and Future Directions

In 2008 Shohamy claimed that:

> Language tests should mediate ideologies and practices in more open, democratic and negotiable ways, and prevent the use of tests as powerful mechanisms capable of imposing draconian policies that have no empirical base. This happens especially when language tests violate diversity, when a false view of language development is being dictated through tests, when language is viewed in isolated ways detached from actual use of multilingual codes in communities, when there are empirical data about the advantage of different accommodations that are being denied. . . (p. 372).

Nevertheless, the Italian situation, in terms of linguistic management of immigration, shows that an "open, democratic, and negotiable way" of viewing

competence in language is still far from being achieved and it opens "old wounds" within reflection on the CEFR: the plurilingualism of its theoretical approach vs. the monolingualism of its descriptors, norm vs. competence as a parameter to describe linguistic competence, and uses and misuses of its levels as political benchmarks.

As Van Avermaet (2010) suggests, there is a certain incongruity between the goals of the CEFR and the profiles of its addressees: "the CEFR descriptors at the lower levels clearly imply an already existing basic knowledge and literacy. [. . .] The CEFR descriptors at higher levels presuppose higher levels of education" (p. 21). These are therefore two conditions which could be problematic if the CEFR descriptors are used to design language programs for integration and as a theoretical model of language tests for low-literate learners. Such situations reflect an idea of linguistic competence as static, mono-normative and artifact (Extra et al. 2009) and show an ideological-linguistic basis hiding behind the CEFR as well.

## Cross-References

- ▶ Assessing English Language Proficiency in the United States
- ▶ High-Stakes Tests as De Facto Language Education Policies
- ▶ Methods of Test Validation
- ▶ Qualitative Methods of Validation
- ▶ Washback, Impact, and Consequences Revisited

## Related Articles in the Encyclopedia of Language and Education

Sally Magnan: The Role of the National Standards in Second/Foreign Language Education. In Volume: Second and Foreign Language Education

Guus Extra: Language Policy and Education in the New Europe. In Volume: Language Policy and Political Issues in Education

Tarja Nikula: CLIL: A European Approach to Bilingual Education. In Volume: Second and Foreign Language Education

## References

Baynham, M., & Moyer, M. (2012). *Language and hyperdiversity in the global city: Re-thinking urban contexts*. Thematic Session at SS19, Berlin, 24–26 Aug 2012.

Blommaert, J., & Rampton, B. (2011). Language and superdiversity. *Diversities, 13*(2), 1–21.

Blommaert, J., Leppänen, S., & Spotti, M. (2012). Endangering multilingualism. In J. Blommaert, S. Leppänen, P. Patha, & T. Raisanen (Eds.), *Dangerous multilingualism* (pp. 1–21). London: Palgrave.

Byram, M., & Parmenter, L. (2012). *The common European framework of reference. The globalization of language education policy*. Bristol/Buffalo/Toronto: Multilingual Matter.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework in simple. *Language Testing, 29*(1), 19–27.
Committee of Ministers. (2008). *Recommendations to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*. Strasbourg: Council of Europe.
Council of Europe. (1996). *The Common European framework of reference for languages: Learning, teaching and assessment*. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf. Accessed 2 Aug 2015.
Council of Europe. (2001). *The Common European framework of reference for languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press. http://www.coe.int/t/dgg/linguistic/source/manual/revision-proofread-final_en.pdf. Accessed 2 Aug 2015
Council of Europe. (2009). *Manual for relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment*.
Council of the European Union. (2002). *Resolution on linguistic diversity and language learning*. Brussels: European Communities.
Cumming, A. (2009). Language assessment in education: Tests, curricula and teaching. *Annual Review for Linguistics, 29*, 90–100.
EC Action Plan. (2004–2006). *Promoting language learning and linguistic diversity*. Brussels: European Communities.
Extra, G., & Yağmur, K. (Eds.). (2012). *Language rich Europe: Trends in policies and practices for multilingualism in Europe*. Cambridge: Cambridge University Press.
Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives*. London: Continuum Press.
Harsh, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11*(2), 152–169.
Kelly, M. (2008). *Hyperdiversity: The challenge of linguistic globalization*. Paper presented at Languages of the Wider World: Valuing Diversity, SOAS, University of London.
Masillo, P. (2015). *Etica e valutazione: uno studio di comparabilità di due test di competenza di livello A2 per adulti immigrati in Italia*. PhD dissertation, University for Foreigners of Siena (Italy).
McNamara, T. (2009). Validity in language testing: The challenge of Sam Messik's legacy. *Language Assessment Quarterly, 3*(1), 31–35.
McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(4), 500–515.
Morrow, K. (Ed.). (2004). *Insights from the common European framework*. Cambridge, UK: Cambridge University Press.
North, B. (2001). *The development of a common framework scale of language proficiency*. New York: Lang.
North, B. (2002). Developing descriptor scales of language proficiency for the cef common reference levels. In J.C. Alderson (Ed.), Common Europian framework of reference for languages: learning, teaching, assessment. case studies. Strasbourg, France: Council of Europe.
Shohamy, E. (2004). Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton & K. Toohey (Eds.), *Critical pedagogies and language learning* (pp. 72–93). New York/London: Cambridge University Press.
Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. New York: Routledge.
Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education, 14*(1), 117–130.
Shohamy, E. (2008). Language policy and language assessment: The relationship. *Current Issues in Language Planning, 9*(3), 363–373.
Spolsky, B. (2008). Introduction: Language testing at 25: Maturity and responsibility? *Language Testing, 12*(3), 321–340.

Trim, S. (2010). The modern languages programme of the Council of Europe as a background to the English Proficiency Project. *English Profile Language Journal, 1*(1), 1–12.
Van Avermaet, P. (2008). *Language policies for the integration of adult migrants: Tendencies in Europe. Some observations and reflections*. Strasbourg: Council of Europe.
Van Avermaet, P. (2010). *Language Requirements for adult migrants: Results of a survey – observations and challenges*. Strasbourg: Council of Europe.
Vertovec, S. (2006). *The emergence of super-diversity in Britain. Centre for Migration, Policy and Society.* Working Paper, 25. Oxford: University of Oxford.
Vertovec, S. (2007). Superdiversity and its implications. *Ethnic and Racial Studies, 30*(6), 1024–1054.

# Assessing English Language Proficiency in the United States

Luis E. Poza and Guadalupe Valdés

**Abstract**

The 2001 reauthorization of the Elementary and Secondary Education Act (ESEA) as the No Child Left Behind Act (NCLB) ushered in a series of important changes for the education of students classified as English language learners, including greater attention to achievement and equity through mandated evaluation and reporting on the part of districts and states of student subgroups. This feature of the ESEA reauthorization and the role of high-stakes testing in general has fueled extensive discussions of educational reform in the years since NCLB, continuing into the present day when the accountability requirements of NCLB are coupled with the benchmarks and assessments set forth in the Common Core State Standards (CCSS). While the CCSS and their accompanying assessments strive to address early criticisms of NCLB such as the narrowing of curriculum, states have also had to develop or adopt new standards for *English as a second language* (most frequently referred to as English Language Proficiency (ELP) Standards). These state ELP Standards are an essential and defining element of the education of English language learners (ELLs) in the context of the CCSS for the foreseeable future and will dictate exactly how learning English is defined for this population, whether a common definition of English language learners can be established, and the degree to which the United States can provide a first-class

L.E. Poza
School of Education and Human Development, University of Colorado Denver, Denver, CO, USA
e-mail: luis.poza@ucdenver.edu

G. Valdés (✉)
Race, Inequality, Language and Education (RILE) Program, Graduate School of Education, Stanford University, Stanford, CA, USA
e-mail: gvaldes@stanford.edu

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_30

education for students who range from emergent to accomplished multicompetent users of two languages.

**Keywords**

English language learners • Assessment • English Language Proficiency Standards

## Contents



## Introduction

Pending implementation of state-level plans for evaluation and accountability under the newly authorized Every Student Succeeds Act (2015), the most recent reauthorization of the Elementary and Secondary Education Act (ESEA) came about in 2001, also known as The No Child Left Behind Act (NCLB). Among its many revisions from previous iterations of the ESEA, originally conceived to provide federal support to students in poverty, was an emphasis on accountability marked by yearly testing and benchmarks for all students to reach proficiency by 2014. While NCLB addresses far more than testing (No Child Left Behind Act (NCLB) 2002), these mandates are the most concrete and tangible for students, teachers, and parents on a daily basis, especially those classified as English language learners (ELL).

In 2009, the Council of Chief State School Officers issued the Common Core State Standards (CCSS), developed with support from researchers and foundations. The new focus on college and career readiness increased the cognitive and linguistic demands compared to most existing state standards. These standards did not affect the accountability requirements of NCLB, but, for states that voluntarily adopted

CCSS to secure additional federal funding for their schools, it did supplant the standards that had been in place. To date, 44 states and the District of Columbia have adopted CCSS, and other states have either adopted the new standards in part or adapted their earlier standards to better align with college and career readiness benchmarks.

In the case of ELLs, a category of students growing in both number and proportion among American public school enrollments (currently, 9.2%, or 4.4 million students (NCES 2015)), both NCLB and CCSS (and potentially ESSA, depending on how states design their English language development trajectories and assessments) create serious equity and opportunity challenges. Under the provisions of NCLB and the new ESSA, states must monitor the academic achievement of ELLs to ensure that they acquire both the English language and the subject-matter competence attained by their English-speaking peers.

## Early Developments

### ELL Classification and NCLB

ELL became an officially recognized category in American federal policy through the 1978 reauthorization of the Elementary and Secondary Education Act, which included provisions for students with "limited English proficiency" (LEP). NCLB likewise provided an explicit and complex definition of the category that includes age, grade level and key student characteristics (e.g., students whose native language is not English, who are born outside the United States, who are Native American or Alaska Native, who come from an environment where a language other than English has had a significant impact on their level of English proficiency, or who come from an environment where a language other than English is dominant). In part D of the definition, the challenges experienced by the types of students who are to be included in the category are described as involving: difficulties in speaking, reading, writing, or understanding English *which are sufficient to prevent them* from achieving on State assessments, successfully achieving in classrooms, or participating fully in society (NCLB 2002). NCLB required states to report on ELLs' achievement separately and to focus efforts on closing the existing educational achievement gap.

NCLB mandated all states to develop a process that screens and identifies children entering American schools as English language learners, classifies then into levels of ELL proficiency, and determines when and whether they can be reclassified as fluent English proficient (FEP). Each state was required (1) to establish or adopt English language proficiency (ELP) standards for all students identified as non-English-background students, (2) to develop an English language proficiency assessment aligned with the state's ELP standards, and (3) to establish criteria that identify when students have met the required level of English proficiency for reclassification as English proficient. These requirements remain under the ESSA,

albeit with increased requirements for statewide standardization of criteria and procedures for ELL classification and redesignation as English-proficient.

## NCLB and Identified Challenges

From the outset, the assessment of ELLs emerged as a key criticism of the No Child Left Behind legislation. Given that schools failing to meet annual benchmarks (known as Adequate Yearly Progress, or AYP) face penalties such as loss of federal funding or forced turnover of school's administration and staff, students' scores on state tests were (and continue to be) a central concern. This issue drew attention to existing achievement gaps, instructional practices, and learning environments that characterize schooling for many ELLs. While some rightfully praise this heightened attention to inequities, the subsequent explanations for disparities were (and are) oversimplified. Gándara and Contreras (2009) note how language proficiency and ethnicity are conflated in discussions of educational achievement for Latino students and this naively places the onus of leveling the playing field entirely onto resolving presumed language barriers. Moreover, the centrality of language and ELP classifications in explaining disparities ignores the heterogeneity of ELLs in terms of nationality and migration(s); home language(s) and linguistic experiences; schooling history; and degrees of bi/multilingual competencies. With these nuances in mind, it is valuable to revisit the assessment regime formalized by NCLB.

Concerns about the usefulness of the data provided by high-stakes tests implemented by the states were discussed in the years of voluntary testing preceding NCLB and in the early years of its implementation. This remained a central issue in the legislation's evaluation going forward (Abedi 2002; Durán 2008; Kopriva 2008; Solano-Flores 2008; Solórzano 2008). These works highlight numerous reliability and validity issues with large-scale assessments, including the development and norming of tests that ignores: (1) the cognitive developmental differences between bilingual and monolingual children, (2) the linguistic characteristics of test items, and (3) the lack of sociocultural relevance of tests normed without ELLs in mind. The government also carried out its own inquiries into the impact of NCLB on ELL students, beginning with a report filed by the Government Accountability Office (GAO 2006). Additionally, a congressional hearing before the Subcommittee on Early Childhood, Elementary, and Secondary Education took place in March 2007 to further investigate matters of teaching practice, teacher preparation and qualification, schools' and districts' familiarity with and ability to implement recommended practices, and the validity and reliability of evaluation methods for the classification of and measurement of achievement among ELLs (Impact of No Child Left Behind on English language learners 2007). Testimony given to Congress, along with the aforementioned GAO report, noted that ELL academic achievement had not improved in accordance with NCLB progress benchmarks in most states, leading to widespread calls for greater flexibility and support from the federal government, as well as efforts to revise existing standards and tests.

## Major Contributions

### The Shift to Common Core State Standards

The CCSS were developed with the ambition of promoting complex thought across disciplines by encouraging students to engage in more analysis, synthesis, and argumentation, as well as to standardize benchmarks across states. This directly addressed two important critiques of NCLB – the isolation and dilution of skills and content as teachers engaged in "teaching to the test" (Gutiérrez et al. 2002; Taylor et al. 2002), along with the inconsistency in standards that made comparing achievement data difficult. The CCSS, however, do not displace any of the accountability or appropriation provisions of NCLB nor the ESSA. Rather, they are merely a new set of standards intended to replace those that states devised independently at NCLB's outset. While the new standards were not mandated, funding made available through the American Recovery and Reinvestment Act of 2009 and the Race to the Top grant program was contingent upon states adopting these standards or devising their own similar in scope and aim.

Many changes have taken place around the country as State Education Agencies (SEAs) and Local Education Agencies (LEAs) moved to implement the new standards. Two consortia, the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC), were funded to develop performance assessments designed to measure the knowledge and skills established by CCSS and were implemented for the first time in 2015. Most importantly, states have also had to develop or adopt new standards for *English as a second language* (most frequently referred to as English Language Proficiency (ELP) Standards). These state ELP Standards are an essential and defining element of the education of ELLs in the context of the CCSS for the foreseeable future and will dictate exactly how learning English is defined for this population.

### Establishing ELP Standards

Given the confusion among states and practitioners about what ELP standards should include in order to correspond to the CCSS and NGSS, in 2010, the Council of Chief State School Officers empaneled a committee of scholars and practitioners to draft a guiding document titled *Framework for English Language Proficiency Development Standards corresponding to the Common Core State Standards and the Next Generation Science Standards* (CCSSO 2012). The document offers direction to states on the development of ELP standards and emphasizes the need for states to clearly articulate and justify their views on language making a clear coherent conceptualization of language and the language acquisition process. The document also states (p. 6) that while it does not support any specific organization of ELP standards:

> The Framework does require that state ELP standards reflect a principled organizational strategy rooted in theoretical foundation(s) that reflects the variety of ways in which different ELLs progress diversely in their language development, including methodologies for scaling and developing descriptions of language proficiency which have been cited and researched. Justification should also be provided for the number of levels adopted and evidence provided to support how these levels represent distinctions that can reasonably be measured and are based on actual student performance.

It is not clear that states and consortia have been guided by this framework. Some states (e.g., Texas, California, New York) have developed new CCSS-aligned ELP Standards, while other states have adopted the standards produced by two different funded consortia: WIDA's Assessment Services Supporting English Learners through Technology Systems [ASSETS] and CCSSO's English Language Proficiency Assessment for the 21st Century [ELPA21). There are many differences between these sets of standards and very dissimilar terms are used in descriptors or performance definitions for establishing the different levels of proficiency. Little justification is provided for the assumed progressions, and limited information is provided about the theoretical foundations that undergird the assumptions made about second language acquisition and development over time.

## Defining and Applying "ELL" as a Category

The adoption of CCSS has also raised awareness of inconsistencies in policy and practice related to the definition of English language learners and their identification. Because of its complexity, there have been many interpretations of the federal definition of English language learners as well as the many differences in the operationalization of the definition by states. The problem of interpretation has been pointed out consistently by a number of researchers over a period of several years (Linquanti 2001; Ragan and Lesaux 2006), and, because of the many questions raised by these inconsistencies, a recent National Research Council study (2011) examined the issue. The panel's report also characterized the ESEA definition as complex and as posing significant problems for the allocation of funds to assist states in serving students determined to be *limited English proficient* (LEP). After examining the GAO (2006) study on data sources available for allocation of funds for ELLs, the NRC report further concluded that no less than *three different definitions* were being employed to identify the LEP/ELL population. Importantly, the panel identified *different conceptualizations of academic and social language* measured by current tests as a significant aspect of the broader problem. It also emphasized that, given these different conceptualizations, state English Language Proficiency (ELP) tests: (1) have different performance levels and (2) test different skills, which are described and measured differently. Because of these differences, students classified at one level (e.g., intermediate) by one state might be classified at an entirely different level in another.

Given pressures brought about by the adoption of the new Common Core State Standards, the question of defining the category of English language learners more

precisely has received increasing attention (e.g., Williams 2014). According to Linquanti and Cook (2013), the US Department of Education has required states participating in any of the four federally funded assessment consortia to adopt a common definition of English learner. As researchers (with the support of the Council of Chief State School Officers) work to inform this process, they report (Linquanti and Cook 2013) that finding a common definition is neither simple nor straightforward. The process will involve four different steps: (1) the identification of potential ELLs, (2) the classification of ELLS in terms of their proficiency levels, (3) the establishment of an English language proficiency criterion against which to assess students, and (4) multiple exit criteria procedure for reclassifying students as fluent English proficient. Williams (2014) contends that the current chaos surrounding the exiting of children from language services can only be remedied by actions at the Federal, State, Assessment Consortia, and District levels working in concert to define and deliver what students actually need in order to succeed in school. For that to occur, policies must be standardized and well-defined.

## Works in Progress

### Supporting an Equity and Opportunity Agenda

The debates about the usefulness of large-scale assessments, setting appropriate standards, and adequately classifying, assessing, and keeping track of ELLs continue. One particularly active collaboration on this front is the Working Group on ELL Policy (http://ellpolicy.org), whose members labor to provide adequate context on the impacts and history of ESEA upon ELLs (Gándara 2015). The group also recommends ways to improve accountability protocols within ESEA through measures such as stabilizing classification protocols such that schools are not penalized for effectively having students reclassify as proficient in English, establishing realistic yet rigorous timelines based in research findings for students to reach acceptable levels of English language proficiency, and setting academic achievement criteria that aligns with students' linguistic proficiencies and language development trajectories (Hopkins et al. 2013).

## Problems and Difficulties

### Getting Language Right

In the case of students categorized as English language learners, every aspect of the educational system that involves them implicates language. Standards, curriculum, pedagogies, and assessments can potentially contribute to or undermine these students' opportunity to develop their subject-matter knowledge. Consequently, it is of vital importance that researchers and practitioners continue to scrutinize the set of progressions and expectations for the development of English language learning

currently mandated by law. Minimally, state systems designed to meet the needs of ELLs must be examined to determine whether they are informed by the body of knowledge (i.e., the scholarship and the research) that is currently available about what language is and how it works, what needs to be acquired, and how instruction can impact the acquisition process.

Views and understandings of language that are established in ELP standards are critical. If they are to serve the purpose of appropriately supporting and monitoring the growth of English language proficiency in ELLs, they must be constructed to describe the trajectory to be followed by K-12 learners in the learning of English based as accurately as possible. Getting this aspect of language right matters because statements about students' expected development contained in ELP standards will establish for parents, for policy makers, for school administrators, and for practitioners:

- The ways that ELL students are assumed to grow in their use of English over time
- The language abilities expected at different levels of development
- The aspects of language that will need to be measured in determining progress
- The types of support that will be required in order to provide these learners with access to instruction in key subject-matter areas (available exclusively in English)

Unfortunately, there is much debate and disagreement surrounding the process of second language acquisition (for a review of early theories and emerging approaches, see Atkinson 2011). There is currently no theoretical consensus about how second languages are acquired, what elements are acquired in what order, whether they can be sequenced and taught, and what needs to be acquired in order for students to use a second language to learn subject-matter content. Educators and members of the public also disagree about what is commonly referred to as *language proficiency*.

The first challenge in establishing state ELL policy and practice systems that can support an equity and opportunity agenda is agreeing on an informed conceptualization of language. Conceptualizations of language are notions and broad ideas about language as well as definitions of language that are informed by the study of or exposure to established bodies of knowledge, by facts about existing and developing theories in applied or theoretical linguistics, by research data on the teaching and learning of second languages, and/or by personal experiences with language and language instruction (Seedhouse et al. 2010).

A second challenge in the development of ELP mandated standards involves establishing an organizational strategy rooted in the knowledge base and scholarship from the field of second language acquisition (SLA) for describing students' developing language proficiencies that includes both a conceptualization of language and an accompanying theory of how language (as conceptualized) is acquired. Obtaining consensus on these issues is difficult, however, because, like many other scholarly fields, SLA is characterized by debates, new perspectives and reexaminations of established views that raise questions about established language-teaching

pedagogies and their underlying theories. For example, within recent years there has been an increasing shift in SLA away from a predominant view of second language (L2) learning/acquisition as an individual, cognitive process that takes place in the mind of individual learners to a view of L2 acquisition as a social process that takes place in interactions between learners and speakers of the target language to be acquired. (Firth and Wagner (1997) is viewed by many as the seminal publication in this turn). Currently, there is increasing agreement on the following points. Second language acquisition is a highly variable and individual process. It is not linear. Ultimate attainment for most L2 learners does not result in monolingual-like language even when the L2 is acquired by very young children (Ortega 2009).

Importantly, for those charged with developing ELP standards documents as well as constructing progressions and stages of language development, existing scholarship reflects much concern about the lack of longitudinal studies in SLA (e.g., Ortega and Iberri-Shea 2005). Researchers working from the tradition of corpus linguistics, for example, argue for authentic collections of learner language as the primary data and the most reliable information about learner's evolving systems. Hasko (2013), drawing from the study of learner corpora, summarizes the state of the field on the "pace and patterns of changes in global and individual developmental trajectories" as follows:

> The amassed body of SLA investigations reveals one fact with absolute clarity: A "typical" L2 developmental profile is an elusive target to portray, as L2 development is not linear or evenly paced and is characterized by complex dynamics of inter- and intralearner variability, fluctuation, plateaus, and breakthroughs. (Hasko 2013, p. 2)

In sum, the state of knowledge about stages of acquisition in L2 learning does not support precise expectations about the sequence of development of English by the group of students whose proficiency must be assessed and determined by the corresponding federally mandated ELP language assessments, and thus, constructing developmental sequences and progressions is very much a minefield. As Larsen-Freeman (1978) argued over 35 years ago, what is needed is an index of development that can serve as a developmental yardstick by which researchers can expediently and reliably gauge a learner's proficiency in a second language broadly conceived.

The third challenge in establishing ELL policies that support equity and opportunity for ELLs is the production of language assessments that correspond to state ELP standards. As pointed out above, ELP Standards establish a conceptualization of language (i.e., what it is that students must acquire). They also describe the order and sequence of the acquisition process so that ELP assessments can then evaluate how well students have learned (or acquired) specific elements, functions, skills, or other aspects of language described in the standards. Assessment is essential for compliance with existing legal mandates.

Assessing language proficiency, however, is a complicated endeavor. As Fulcher and Davidson (2007, p. 2) contend, the practice of language testing "makes an

assumption that knowledge, skills and abilities are stable and can be 'measured' or 'assessed.' It does it in full knowledge that there is error and uncertainty, and wishes to make the extent of the error and uncertainty transparent." Importantly, there has been an increasing concern within the language testing profession about the degree to which that uncertainty is actually made transparent to test users at all levels as well as the general public. Shohamy (2001), for example, has raised a number of important issues about ethics and fairness of language testing with reference to language policy. Attention has been given, in particular, to the impact of high-stakes tests, to the uses of language tests for the management of language-related issues in many national settings, and to the special challenges of standards-based testing (Cumming 2008). Cumming (2008, p. 10.), for example, makes the following very strong statement about the conceptual foundations of language assessments:

> A major dilemma for comprehensive assessments of oracy and literacy are the conceptual foundations on which to base such assessments. On the one hand, each language assessment asserts, at least implicitly, a certain conceptualization of language and of language acquisition by stipulating a normative sequence in which people are expected to gain language proficiency with respect to the content and methods of the test. *On the other hand, there is no universally agreed upon theory of language or of language acquisition nor any systematic means of accounting for the great variation in which people need, use, and acquire oral and literate language abilities*. (Emphasis added)

Cumming argues that, given this dilemma, educational systems nevertheless develop their own sets of standards through a policy-making consensus process generally based on the professional perspectives of educators or on the personal experiences and views of other members of standards-writing committees rather than empirical evidence or SLA theories. Cumming further points out that this approach involves a logical circularity because what learners are expected to learn is defined by the standards, taught or studied in curriculum, and then assessed "in reference to the standards, as a kind of achievement testing." (p. 10)

According to Cumming, then, ELP assessments, as currently constructed, tell us very little about students' proficiency or competency in English broadly conceived. They can only tell us where a student scores with reference to the hypothesized sequence of development on which the state assessment is based. Such scores are useful because given current federal and state regulations, they allow educators to classify and categorize students and, in theory, to provide them with instructional supports appropriate for them while they acquire English. Many would argue that in a world of imperfect systems, states are doing the very best they can.

## Language, Opportunity, and Equity

In order to achieve both equity and opportunity for all students, public officials, school administrators, researchers, and educators must begin with a clear

understanding that definitions and categorizations established by federal and state laws, policies, and guidance documents as well as by standards-setting processes arrived at by political consensus may have unintended and serious negative consequences for students. As pointed out above, a recent National Research Council study, *Allocating Federal Funds for State Programs for English Language Learners* (National Research Council 2011), added to our knowledge about these issues. After undertaking the examination of the English Language Proficiency (ELP) assessments currently used by the states, the report concluded that:

> For this set of tests, we found evidence that the assessments have been developed according to accepted measurement practices. Each of the testing programs documented its efforts to evaluate the extent to which the test scores are valid for the purpose of measuring students' language proficiency in English. The tests are all standards-based. They all measure *some operationalized conceptualization of academic language*, in addition to *social/conversational language*, in *four broad domains* and report scores for each of these domains, as well as a comprehension score and one or more composite scores. They all summarize performance using proficiency or performance levels, and states have established methods of looking at overall and domain scores in order to determine *their respective definitions of English language proficiency*. The tests also have versions available for students in kindergarten through 12th grade, with linkages to enable measurement of growth across adjacent grade bands. These common features provide the foundation for a certain degree of comparability across the tests. (NRC 2011, p. 74. Emphasis added)

As will be noted, the panel identified *different conceptualizations of academic and social language* measured by current tests but focused on the fact that distinguishing between academic and social language was common across the assessments analyzed. It did not problematize or compare these various perspectives, but it did note that the definition of proficiency is determined differently in each state. The panel pointed out, moreover, that tests have different numbers of performance levels, test different skills which are themselves described and measured differently, and that students classified at one level (e.g., intermediate) by one state might be classified at an entirely different level in another. The panel considers several different methods that might be used to establish comparability but concludes by stating that cross-state comparability was not a goal in the development efforts of existing ELP assessments.

## Future Directions

There are several key areas to prioritize in the process of improving assessments and the accountability systems they underlie to make them more equitable for ELLs. One is a more consistent and realizable definition of the ELL label itself across states and districts. This requires more uniform protocols to screen students as they enter schools for initial classification, careful attention to avoid misclassification of students into Special Education simply over language issues, and also for their eventual reclassification as proficient in English. In moving toward a common definition of

English language learners (Linquanti and Cook 2013), it is evident that both conceptualizations of language and theories of the ways in which language is acquired matter. If we are to develop a "common performance level descriptor" (PLD) for "English proficient" as advocated by Linquanti and Cook (2013), such a descriptor cannot be based on a political consensus that results in contradictory or incompatible conceptualizations of language or on descriptions and progressions of language acquisition that are not informed by the currently shifting knowledge about the process of acquisition in the field of SLA. In order to develop a common performance level descriptor, we must engage in the task of defining the ways that proficiency can be conceived from various theoretical perspectives. We must weigh the alternatives, argue about contradictory positions, and consider the pedagogical implications of these alternatives. To be sure, the process of defining and conceptualizing language in the light of academic debates about both language and second language acquisitions will be complex, time-consuming, and expensive, but it can and must be engaged.

Further, ongoing work on improving the schooling experiences and outcomes of ELL students must further attend to the heterogeneity in the ELL population rather than be contented with oversimplified "language barrier" explanations for disparities as is often the case among practitioners and policymakers now. In this vein, the emergent scholarship on language as a social practice and evolving repertoire of skills and features must add to its thorough and valuable qualitative descriptions of learning and meaning-making in classroom interactions some evidence of systemic improvement if these principles are to translate into common pedagogical practice. Getting language right for such purposes is an enormous challenge. The stakes, however, have never been higher. The United States cannot afford to provide a second-class education to its growing number of English language learners (Gándara & Orfield 2012), whether as part of the current educational reform movement or as part of a plan for the future of the nation.

This chapter was submitted for review prior to the authorization of the Every Student Succeeds Act. Nevertheless, many of the stated challenges persist. ESSA and subsequent regulations call for states to create statewide, uniform objective criteria for classifying, evaluating, and measuring progress of ELL students toward proficiency within a state-determined time frame. States must also account for student characteristics such as initial English proficiency when determining English proficiency targets. Most notably, ESSA moves accountability for ELL progress into Title I (from Title III under NCLB), which is the primary lever of school accountability attached to a much larger pool of federal funds. These changes help draw attention to the education of ELL students, recognize their heterogeneity, and standardize criteria for classification into and redesignation from EL status. However, concerns remain regarding variability across states; quality of instruction, assessment, and curriculum for EL classified students; and the setting of appropriate targets to determine proficiency informed by the latest research on bilingualism and bilingual language development.

## Cross-References

- ▶ The Common European Framework of Reference (CEFR)
- ▶ Utilizing Accommodations in Assessment
- ▶ Washback, Impact, and Consequences Revisited

## Related Articles in the Encyclopedia of Language and Education

Wayne Wright, Thomas Ricento: Language Policy and Education in the USA. In Volume: Language Policy and Political Issues in Education

Katherine Schultz, Glynda Hull: Literacies In and Out of School in the United States. In Volume: Literacies and Language Education

Patricia Gandara, Kathy Escamilla: Bilingual Education in the United States. In Volume: Bilingual and Multilingual Education

## References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*(3), 231–257.

Atkinson, D. (Ed.). (2011). *Alternative approaches to second language acquisition*. New York: Routledge.

Council of Chief State School Officers (CCSSO). (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. http://www.ccsso.org/Documents/2012/ELPD%20Framework%20Booklet-Final%20for%20web.pdf. Last accessed 16 June 2015.

Cumming, A. (2008). Assessing oral and literate abilities. In *Encyclopedia of language and education* (pp. 3–17). New York, NY: Springer.

U.S. Department of Education, National Center for Education Statistics. (2015). The condition of education 2015: English language learners (NCES 2015–144). https://nces.ed.gov/programs/coe/indicator_cgf.asp. Last accessed 27 June 2015.

Durán, R. (2008). Assessing English-language learners' achievement. *Review of Research in Education, 32*(1), 292–327.

Every Student Succeeds Act. Pub. L. No. 114–95. 114th Congress. (2015–2016)

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *Modern Language Journal, 81*, 285–300.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London/New York: Routledge.

Gándara, P. (2015). *Charting the relationship of English learners and the ESEA: One step forward, two steps back*. RSF: The Russel Sage Foundation Journal of the Social Sciences. 1(3). 112–128.

Gándara, P., & Contreras, F. (2009). *The Latino education crisis: The consequences of failed social policies*. Cambridge, MA/London: Harvard University Press.

Gándara, P., & Orfield, G. (2012). Segregating Arizona's English learners: A return to the "Mexican room". *Teachers College Record, 114*(9), 1–27.

Government Accountability Office (GAO). (2006). *No Child Left Behind Act*: *Assistance from Education could help states better measure progress with limited English proficiency.* http://www.gao.gov/products/GAO-06-815. Last accessed 16 June 2015.

Gutiérrez, K. D., Asato, J., Santos, M., & Gotanda, N. (2002). Backlash pedagogy: Language and culture and the politics of reform. *The Review of Education, Pedagogy & Cultural Studies, 24*(4), 335–351.

Hasko, V. (2013). Capturing the dynamics of second language development via learner corpus research: A very long engagement. *The Modern Language Journal, 97*(S1), 1–10.

Hopkins, M., Thompson, K. D., Linquanti, R., Hakuta, K., & August, D. (2013). Fully accounting for English learner performance a key issue in ESEA reauthorization. *Educational Researcher, 42*(2), 101–108.

Impact of No Child Left Behind on English language learners: Hearing before the Subcommittee on Early Childhood, Elementary, and Secondary Education, Committee on Education and Labor. US House of Representatives, 110th Congress, 1 (2007). http://www.gpo.gov/fdsys/pkg/CHRG-110hhrg34017/pdf/CHRG-110hhrg34017.pdf. Last accessed 16 June 2015.

Kopriva, R. (2008). *Improving testing for English language learners*. New York: Routledge.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*(4), 439–448.

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners*. University of California Linguistic Minority Research Institute.

Linquanti, R., & Cook, H. G. (2013). *Toward a "common definition of English learner": Guidance for states and state assessment consortia in defining and addressing policy and technical issues and options*. Washington, DC: Council of Chief State School Officers.

National Research Council (NRC). (2011). *Allocating federal funds for state programs for English language learners*. Washington, DC: National Academies Press.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, § 115. Stat, 1425 (2002).

Ortega, L. (2009). Sequences and processes in language learning. In H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 81–105). Malden: Blackwell.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26–45.

Ragan, A., & Lesaux, N. (2006). Federal, state, and district level English language learner program entry and exit requirements: Effects on the education of language minority learners. *Education Policy Analysis Archives, 14*(20). doi:10.14507/epaa.v14n20.2006.

Seedhouse, P., Walsh, S., & Jenks, C. (2010). *Conceptualising 'learning' in applied linguistics*. New York: Palgrave Macmillan.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Longman.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*(4), 189–199.

Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78*(2), 260–329.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2002). A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost. CSE Technical Report 588. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles: University of California.

Williams, C. P. (2014). *Chaos for dual language learners: An examination of state policies for exiting children from language services in the PreK-3rd grades*. New America. http://www.newamerica.org.s3-website-us-east-1.amazonaws.com/downloads/chaosfordlls-conorwilliams-20140925_v3.pdf. Last accessed 16 June 2015.

# Critical Language Testing

Elana Shohamy

**Abstract**

Critical language testing (CLT) refers to the examination of the uses and consequences of tests in education and society (Shohamy 2001a, b; Spolsky 1995). The topic gained attention by various scholars and particularly Messick (1981, 1989), who argued for expanding the definition of construct validity as a criterion for evaluating the quality of tests, to include components related to tests use, such as values, impact, and consequences. CLT emerged from the realization that tests are powerful tools in education and society, which may lead to unintended consequences that need to be examined and evaluated. It is the power of tests, especially those of high stakes, that causes test takers and educational systems to change their educational behaviors and strategies as they strive to succeed in tests given their detrimental impact.

Ample research on CLT exists which focuses mainly on the uses of tests with regard to high-stakes tests such as the TOEFL, school leaving exams, entrance and placement tests, as well as international/comparative tests such as PISA and TIMMS. These studies pointed to the misuses of tests and their impact that goes far beyond learning and teaching into issues of identity, educational policies, as well as marginalization and discrimination against immigrants and minority groups. The chapter ends with a discussion of alternative testing strategies, developed over the past decade, which aim at minimizing the power and negative consequences of tests mostly by including democratic approaches of formative and dynamic assessment, multilingual testing, inclusive assessment, and bottom-up testing policies and tasks, all aiming to use tests in constructive and positive ways, diminishing their excessive power.

E. Shohamy (✉)
School of Education, Tel Aviv University, Tel Aviv, Israel
e-mail: elena@post.tau.ac.il

© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_26

**Keywords**
Consequential validity • Washback • Multilingual assessment • Ethicality • Democratic assessment • Values • Assessment literacy

## Contents



## Introduction and Early Developments

In most countries worldwide, individuals are subject to tests, whether to enter educational programs, to pass from one level to the next, or to be granted certificates to practice professions. Tests determine whether students will be allowed to enter high schools and higher education and in many cases even kindergartens and elementary schools. In schools, classroom tests are used in all subjects and grades and have an effect on students' status in their classrooms as well as on their identities and self-concepts. Tests are used by teachers as disciplinary tools to control students' behaviors and the curricula and to upgrade the status and prestige of specific topics and subjects. High-stakes tests lead to rejections and acceptances, to winners and losers, and to successes and failures and hence have an impact on people's lives. For adult immigrants, tests determine whether they will be granted permission to immigrate and to obtain citizenship in countries they moved to or seek asylum.

Critical language testing (CLT) originated from a focus on the *uses* of language tests and the realization of their enormous power to influence education, societies, and even the status of nations as a result of performances on international tests. It is the power of tests and the detrimental decisions they bring about that grants them such status in society so that people change their behavior in order to succeed on tests (Shohamy 2001a). It is this very power that brings about decision makers and those in authority to introduce tests since they know that once a high-stakes test is introduced, it is most likely that principals, even if the curriculum has not changed, will start imposing the teaching of these topics, and students will be forced to learn them. Hence, there is a change in stakeholders' behaviors in an intensive effort to achieve high scores. In fact, in many schools the content that is included in these tests becomes the de facto curriculum and often overlooks the written curriculum that already exists as those who introduce tests often have different educational agendas (Cheng 2004; Cheng and Curtis 2004 and others).

Two examples that demonstrate the phenomenon are the following: The first in the context of migration (Extra et al. 2009), where adult immigrants, moving

to a new country are required to take tests of their proficiency in the language used in the new location as a condition for citizenship and residence. At times these tests are being administered still in their home countries and thus restrict the number of immigrants. Governments implement language testing regimes as a way to control the number of immigrants they allow to enter the country and/or of those who can stay there. In most nations nowadays immigrants are required to pass a test in the main official language of the country. This policy does not originate from research findings that demonstrates that proficiency in national languages is relevant for functionality; still, language tests become the tool for screening, leading to decisions as to whether immigrants are allowed to stay in the country or would be forced to leave. It also ignores situations when immigrants are at an age that they are incapable of learning the new language and/or cannot read or write in their own language or when there are no learning opportunities such as language courses where they can learn the new language (McNamara and Shohamy 2008; Shohamy and Kanza 2009). It is also known that many immigrants tend to be employed in their own communities and are very comfortable using their home languages which are functional for them in most domains of everyday lives. The test then is used primarily as a tool to screen immigrants, which brings about enormous criticism about the ethicality of these types of tests as they are used for purposes they were not intended to. The children of immigrants usually acquire the new language relatively fast in comparison to their parents because they are schooled in that new language as a medium of instruction, albeit, this too takes a long time (Levin and Shohamy 2008) as will be reported below.

The second case is the testing of immigrant and minority school students who lack high proficiency in the power language which is the medium of instruction in schools. In this case students are required to take standardized tests as mandated by national policies after a short time of being in the country. While research shows that it takes immigrants about 10 years to acquire a new language (Collier and Thomas 2002; Valdés et al. 2015; Levin and Shohamy 2008) and yet while they are still in the process of learning the new language, they are being tested in school content areas via the new language. Given that the students are not proficient in the language yet, they often fail these tests in the different academic subjects and become marginalized and discriminated against by their teachers and peers (Levin and Shohamy 2008; Levin et al. 2003).

In both of the above-described cases, language testing policies are used as disciplinary tools given that test takers have no choice but to comply with the policy demands. While test takers and regional educational systems comply with such disciplinary demands, they also resent them as they feel they were imposed on them without their voice being heard. It is the powerful uses of tests – their detrimental effects and their uses as disciplinary tools that are responsible for the strong feelings that tests evoke in test takers. It is the raising of critical questions about the testing policy and their impact and consequences as well as the intentions behind the introduction of these tests which is the essence of CLT.

## Major Contributions

**A social perspective**. The use of tests for power and control was argued convincingly by Foucault. In *Discipline and Punish: The Birth of the Prison* (1979) Foucault stated that examinations possess built-in features that enable them to be used for exercising power and control. Specifically he mentions that tests serve as means for maintaining hierarchies and normalizing judgment. They can be used for surveillance, to quantify, classify, and punish. Their power lies in that they can lead to differentiation among people and for judging them. Tests consist of rituals and ceremonies along with the establishment of truth and all in the name of objectivity, as Foucault puts it:

> The examination combines the techniques of an **observing hierarchy** and those of a **normalizing judgement**. It is a **normalizing** gaze, a **surveillance** that makes it possible to **qualify,** to **classify** and to **punish**. It establishes over individuals a **visibility** through which one **differentiates** them and **judges** them. That is why, in all the mechanisms of discipline, the examination is highly **ritualized**. In it are combined the **ceremony of power** and the **form of the experiment**, the **deployment of force** and the **establishment of truth**. At the heart of the procedures of discipline, it **manifests the subjection** of those who are perceived as objects and the **objectification** of those who are subjected. (p. 184) (my emphasis)

In Foucault's biography, written by Eribon (1992), he provides evidence of Foucault's personal experiences and sufferings from tests, making him a "test victim." He shows that Foucault himself was a victim of tests, who failed on high-stakes tests. References are made to situations when tests played detrimental roles in his own life, possibly causing him to gain the special insight into the uses of tests as disciplinary tools. Foucault (1979) also noted that it is only in the twentieth century that testers made tests "objective unobtrusive" messengers, while in the past testers had to face test takers directly and to share the responsibility for the testing verdict.

The notion that tests represent a social technology is introduced by Madaus (1990) as an extension of the uses of tests as disciplinary tools. He claimed that tests are scientifically created tools that have been historically used as mechanisms for control and their power is deeply embedded in education, government, and business. The test is a means for social technology as it not only imposes behaviors on individuals and groups but also defines what students are expected to learn and know and can therefore be referred to as "de facto curriculum." It therefore guaranteed the movement of knowledge from the teacher to the pupil, but it extracted from the pupil a knowledge destined and reserved for the teacher.

Bourdieu (1991) claimed that tests serve the needs of certain groups in society to perpetuate their power and dominance; thus, tests were rarely challenged. Tests have wide support of parents, as they lead to the imposition of social order. For parents who often do not trust schools and teachers, tests provide indication of control and order, especially given their familiarity with tests in their own years of schooling. For many parents tests symbolize control and discipline and are perceived as indications of effective learning. It is often observed that raising the educational standards

through testing appeals to the middle classes, partly as it means gaining access to better jobs for their children, and for some it is also a code word for restricting minority access. The paradox is that low-status parents, minorities, and immigrants, who are constantly excluded by tests, have an overwhelming respect for them and often fight against their abandonment.

Hanson (1993) as well discusses the power of tests to affect and define people and notes that tests have become social institutions on their own, taken for granted with no challenging questions. Specifically, while a testing event is only a minute representation of the whole person, tests are used both to define and predict a person's ability as well as to keep them powerless and often under surveillance. He adds the following:

> In nearly all cases test givers are (or represent) organizations, while test takers are individuals. Moreover, test-giving agencies use tests for the purpose of making decisions or taking actions with reference to test takers – if they are to pass a course, receive a driver's license, be admitted to college, receive a fellowship, get a job or promotion. . . That, together with the fact that organizations are more powerful than individuals, means that the testing situation nearly always places test givers in a position of power over test takers. (Hanson 1993, p. 19)

The use of language tests as disciplinary tools by powerful political institutions is discussed by McNamara (1998) who notes that tests have become an arm of policy reform in education and vocational training as well as in immigration policies. Such policy initiatives are seen within the educational systems as well as in the workforce. A concern for national standards of educational achievement in a competitive global economy, together with a heightened demand for accountability of government expenditures, has propelled a number of initiatives involving assessment as an arm of government educational policy in the national, state, and district levels.

**A psychometric perspective.** Some psychometricians who themselves develop tests have been critical about them. Most notable is Messick, who was employed at the Educational Testing Service in the USA, a center that develops and researches tests. Messick (1981, 1996) was among those who drew attention to the topic of impact, claiming that tests' consequences should be incorporated into a broader perspective of a unified concept of validity. He argued that given that social values were associated with intended and unintended outcomes, the interpretations and uses which derive from test scores, the appraisal of the social consequences of tests should be subsumed as aspects of construct validity (1996, p. 13). Messick (1996) claimed that "[i]n the context of unified validity, evidence of washback is an instance of the consequential aspect of construct validity." Thus, Messick's concept of unified validity seems to be the bridge between the narrow range of effects included in washback and the broader one encompassed by "impact" which includes ". . .evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use. . . especially those associated with bias. . . unfairness in test use, and with positive or negative washback effects on teaching and learning" (p. 12). The term "consequences" is used mostly by Messick to encompass washback and construct validity but with a stronger focus on ideological values. This

is also how the term is used here to discuss the societal influences of tests in a larger scope. Messick notes that washback is only one form of testing consequence that needs to be weighed when evaluating validity, and testing consequences are only one aspect of construct validity, leading to the term *consequential validity.* An additional term often used to refer to the connection between testing and instruction is *systemic validity* (Frederiksen and Collins 1989) relating to the introduction of tests into the educational system, along with additional variables which are part of the learning and instructional system. In such situations tests become part of a dynamic process in which changes in the educational system take place according to feedback obtained from the test. Similar terms associated with the impact of tests on learning are *measurement-driven instruction*, referring to the notion that tests drive learning, and *curriculum alignment*, implying that the curriculum is modified according to test results.

Thus, although psychometricians developed sophisticated methods for test development and design, in terms of reliability and validity and quality of items and tasks, they tend to overlook the important dimension of consequences of tests. This leads to the need to pose questions that will incorporate the consequences such as: What are the tests being used for? What purposes are they intended for? Do they lead to decisions which are beneficial or harmful for people? Are they meant to evaluate the level of language proficiency or as sanctions for discipline and control? In other words, is a test really a pure measurement of language proficiency or is it used as a disciplinary tool for other agendas such as selection, expulsion, and differentiation leading to stigmas about different populations and their rejection from bastions of society?

**Language testing perspective.** In the book *Measured Words*, Spolsky (1995) surveyed the different language tests from a historical context. He brings up the cases of the different agendas that were associated with the TOEFL tests to prevent people from certain areas in the world to study in US educational institutions.

Shohamy (2001a) introduced the notion of CLT and focused on three studies which demonstrated how the introduction of high-stakes language tests brought about major changes in the behavior of the school: a test of oral proficiency in English as a second language which led to teaching to the test, a test in Arabic which turned the classes to prepare students for the test which meant studying the exact content of the test, and a national test for testing reading comprehension which in a year's time changed drastically the reading curriculum to texts with multiple questions; in all cases, there was narrowing of the curriculum and the test dominated most activities. In other words, once the test was administered, the teaching returned to non-testing activities. A study (Shohamy et al. 1996) examined the effect of these tests several years later and showed that the only meaningful change took place in the high-stakes tests while in the case of low-stakes tests these changes were totally overlooked. Shohamy and McNamara (2009) critiqued the tests for citizenship. In Shohamy (2009) there is a strong argument and a list of reasons against the use of tests for enforcing such policies.

Studies by Alderson and Wall 1993 examined a number of hypotheses whereby one could expect a change in the school learning policy of languages due to tests but found very little effect due to the tests.

Cheng (2004; Cheng et al. 2011; Cheng and Curtis 2004; Cheng and DeLuca 2011) conducted studies focusing on the washback of high-stakes tests, especially in China but also in Canada and elsewhere. They found major impact of tests on teaching. More information about these studies can be found in the chapter "► Washback, Impact, and Consequences Revisited" by Tsagari and Cheng in this volume, examining at least two different types of washback studies, one related to traditional standardized tests and the other in which modified versions of tests are examined as means for achieving more positive influence on teaching and learning.

Fulcher (2004) critiqued the growing number of rating scales which are expected to provide more accurate scores. Two of these well-known scales are the ACTFL scale used in the USA and the CEFR used mostly in Europe and elsewhere. Yet, major critiques have emerged from these scales, as to their linearity, and the scales not being appropriate for all learning settings. (see a chapter "► The Common European Framework of Reference (CEFR)" by Barni and Salvati, in this volume).

Shohamy's (2001a) brought up pleas for developing more democratic views of tests, increasing the responsibility of testers, minimizing the power of tests, protecting test takers, and posing a questions about the ethical roles of language testers.

One immediate outcome of the CLT was the development of a *Code of Practice* to protect test takers. Davies (this volume), who was very attentive to the notion of CLT, examined the professionalism of language testers who design tests and overlook their impacts. He posed ample questions about what it means to be an ethical and professional tester and their responsibilities. Davies served as the chair of the ILTA (International Language Testing Association) committee that developed the *Code of Ethics* and a *Code of Practice* to be used by language testers in the development and uses of tests, so testers become aware of their professional and ethical responsibilities. The real aim accordingly was to create tests which are more fair, considerate, constructive, and ethical in terms of their power.

## Work in Progress

Over the years, a large number of questions emerged that have fallen under the paradigm of CLT. With the introduction of language citizenship tests for immigrants in an expanding number of countries in Europe, Asia, and elsewhere, ample studies pointed to the harmful effects of those tests. Milani (2008), based on protocols about integration in the Swedish Parliament, pointed to the debates about the tests revealing a taste of discrimination, given the goal of integrating immigrants into the main society. A special issue of the journal *Language Assessment Quarterly* (Shohamy and McNamara 2009) focused on these tests in a number of countries such as Estonia, Latvia, the UK, the USA, and Israel. A number of comprehensive edited books (Stevenson 2009; Extra et al. 2009) were published as well. Unfortunately, this research did not yield major changes in terms of government policies, and the

problem continues as more countries adopt these tests. Recently, Norway is joining these countries with the introduction of new citizenship tests as of January 2017. These policies get stricter as the wave of immigration expands in Europe and elsewhere. Likewise in schools, while tests such as the ones mandated by the NCLB act ceased to exist in the USA, the new policy of the Common Core represents a new testing policy with higher cognitive demands introduced in schools, thus creating injustices for immigrants (Abedi 2001, 2004; Abedi and Dietal 2004; Shohamy and Menken 2015). Thus, these tests, as Valdés and Poza point out, discriminate against newcomers and minority groups (see their chapter "► Assessing English Language Proficiency in the United States" in the present volume).

At the same time, the work on CLT continues (see the chapter "► Washback, Impact, and Consequences Revisited" by Tsagari and Cheng in this volume). Indeed, the notion of the "power of tests" puts enormous responsibility on the shoulders of those who wield the tests. Yet, at the same time there are also new approaches that attempt to respond to the power of tests, to minimize and challenge it by focusing on tests geared for more effective learning rather than tools for punishment. Further, with the changes toward multilingualism, there is more of an emphasis on the meaning and essence of language in this day and age with regard to globalization, multilingualism, language varieties, and mixture of languages to include immigrants and minority groups in different types of multilingual tests.

These directions responded to questions such as the following:

- Do the tests reflect the bi-/multilingual uses of language in this day and age in the context of plurilingual societies?
- Do tests have realistic goals in terms of their levels of proficiency, considering the dynamic and fluid nature of language?
- Are the validation procedures based on realistic norms and not on the native speaker?
- Do they consider all components that contribute to performance, beyond language per se?
- Are we ethical when we design tests based on definitions and goals provided by central agencies?
- Are language tests open to monitoring by society, critiqued, and sanctioned?
- How can immigrants and minority groups be included in spite of their language proficiency, given that they are educated, talented, good people, but have difficulties with language, or it takes them long time to acquire it?
- How can immigrants who have to pay big amounts of money for language courses get resources that will help them learn the languages? And is the "almost" native speaker realistic for all people, regardless of age, background, etc. (see the chapter "► Assessing English as a Lingua Franca" by Jenkins and Leung, which discusses the ELF variety that most English nonnative speakers use, as well as the chapter "► High-Stakes Tests as De Facto Language Education Policies" by Menken, both in this volume)?
- Do all immigrants need to pass a language test where there is no evidence that knowledge of "the" language necessarily contributes to good citizenship? That is,

how can we minimize the use of language tests, preventing them from being a major tool in creating immigration policy?

In the section below, a list of additional new initiatives which can defy or minimize the power of tests will be briefly described.

## Responses: Minimizing and Resisting Power

The topics below include strategies of assessment initiatives and practices which can lead to more positive outcomes of tests which are more fair, just, and mostly educational. These go beyond standardized tests into language testing that proposes means for diverting tests to learning and less for judgment.

**Dynamic assessment**: An approach whereby testing and teaching are connected and hence minimize the power of tests, based on Vygotsky's sociocultural theory whereby the emphasis is the use of tests for learning (see the chapter "► Dynamic Assessment" by Poehner, Davin, and Lantolf in this volume; and see also Levi 2015; and Levi 2016).

**Assessment literacy** increases the basic knowledge about assessment including CLT and focuses on their consequences and impact, which needs to be addressed and is part of language testing along with other factors (see chapters "► Language Assessment Literacy" and "► Training in Language Assessment" by Inbar-Lourie and Malone in this volume).

**Test accommodation and differential item functioning (DIF)** provide a tool for assisting immigrants and minority groups who are not familiar with the new language to obtain assistance and thus enhance the achievement in academic and content subjects, especially for the early years of migration while learning the new language (see chapter "► Utilizing Accommodations in Assessment" by Abedi in this volume). Further, the focus is on the technique of DIF as a strategy to identify the test items and tasks which discriminate against students of different backgrounds. Removal of such items and tasks results in tests which are more fair to larger pool of test takers.

**Formative/alternative assessment** attempts to develop assessment strategies which are more constructive than standard external items, often developed by local agents at the schools and not by central agencies (see the chapter "► Task and Performance-Based Assessment" by Wigglesworth and Frost and the chapter "► Using Portfolios for Assessment/Alternative Assessment" by Fox, in this volume).

**Multilingual/translanguaging and ELF tests**. An approach built on the nature of the language construct as it is being viewed today, where languages are mixed and people use them in very creative ways. Shohamy (2011) demonstrated how the use of multilingual tests in testing mathematics of immigrant students (in Hebrew and Russian on the same test) result in higher mathematics scores than of those students who were tested in monolingual (Hebrew) tests in Israel. A case in point can best be demonstrated with English away from the concept of the native speaker (see the chapter "► Assessing Multilingual Competence" by Lopez, Turkan, and Guzman-Orth and also the chapter "► Assessing English as a Lingua Franca" by Jenkins and Leung in this volume).

**Tests for indigenous contexts.** Given that the essence of testing is that it grants importance to languages and provides a message that they should be empowered, there is a call to include indigenous language within the repertoire of testing (see the chapter "► Language Assessment in Indigenous Contexts in Australia and Canada" by Wigglesworth and Baker in this volume).

**Full language repertoire (FLR).** This refers to expansion of assessment to include all the languages a person knows, regardless of each language level of proficiency. This is especially relevant with immigrant students who arrive in a new location, so the languages they know from the past will not be overlooked and ignored, but rather they should be incorporated into the whole language repertoire, viewing these languages as significant resources.

Other themes included in this volume that have the potential to reduce the power of tests and focus more on learning include the following: the chapter "► Assessing Students' Content Knowledge and Language Proficiency" by Llosa recommending a focus on content and less on language proficiency, the chapter "► Culture and Language Assessment" by Scarino with emphasis on culture within assessment, and qualitative methods of validation by Lazaraton. Chapter "► Assessing the Language of Young Learners" by Bailey especially warns about the overuse of tests with regard to young leaners. Other studies demonstrate the extent to which language tests are instrumental for control. Tsagari and Cheng show how significant it is to examine the consequences of tests so to limit their powerful status, which is related to the chapter "► High-Stakes Tests as De Facto Language Education Policies" by Menken demonstrating that tests should avoid dictating the curriculum but rather reflect it. These are manifested in the use of tests as "de facto" curriculum, approaches which should be minimized. All these are warning signs regarding the ethics, professionalism, rights, and codes as described in the chapter "► Ethics, Professionalism, Rights, and Codes" by the late Alan Davies. The existence of the Common European Framework of Reference (CEFR) requires testers to be even more cautious about the power of tests as these scales provide an extra tool to bring about homogeneity, as can be seen in the chapter "► The Common European Framework of Reference (CEFR)" by Barni and Salvati. The dangers of using tests which are not appropriate to specific students are being critiqued by Poza and Valdés (chapter "► Assessing English Language Proficiency in the United States").

All in all, many of the chapters in this volume discuss and propose a number of ways to focus on learning and hence to minimize the power of tests by using the strategies described above and in many of the chapters.

## Future Directions

CLT led to ample questions about the quality of tests, their consequences, and the difficulties they impose on test takers and systems. Tests often offer simplistic solutions for complex issues. The research in this field attempted to explore areas where tests are misused by examining their consequences and the intentions of those who introduced them. The responses are varied so that what is considered negative

or positive is constantly debated given Messick's views that these are related to the values of test takers, educational systems of nations, and ideologies of governments and regimes to use tests for power and control. Yet, the obligation of those engaged in language testing is to adopt CLT approaches to try to look beyond the tests themselves and toward their uses; in other words, a good test may be necessary but not sufficient. It is the obligation of all those working in test development and use to constantly ask questions as to intentions and uses of tests in education and society with regard to the multiple groups for whom national languages are second languages.

It is encouraging to see that in the past decade a number of different types of assessment strategies and procedures have been developed and implemented. These strategies are currently being used to broaden the construct of testing tests and provide successful ways of "talking back" to the power of tests that can minimize their power and protect test takers and parents, teachers, and principals, enhancing the uses of assessment procedures to minimize discrimination and marginalization and maximize learning, fairness, ethicality, equality, and justice. The purpose is not to eliminate tests but rather to see the values behind them as well as their hidden agendas in the area of accountability and the learning of languages and to reflect perspectives of languages in this day and age.

## Cross-References


- ▶ Assessing English as a Lingua Franca
- ▶ Assessing Multilingual Competence
- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ History of Language Testing
- ▶ High-Stakes Tests as de facto Language Education Policies


## Related Articles in the Encyclopedia of Language and Education


Linda von Hoene: The Professional Development of Foreign Language Instructors in Postsecondary Education. In Volume: Second and Foreign Language Education

Bonny Norton, Ron David: Identity, Language Learning and Critical Pedagogies in Digital Times. In Volume: Language Awareness and Multilingualism

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

Hilary Janks, Rebecca Rogers, Katherine O'Daniels: Language and Power in the Classroom. In Volume: Language Policy and Political Issues in Education

Stephen May: Language Education, Pluralism, and Citizenship. In Volume: Language Policy and Political Issues in Education

## References


Abedi, J. (2001). *Assessment and accommodations for English language learners: Issues and recommendations* (CRESST Policy Brief 4). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4–14.

Abedi, J., & Dietal, R. (2004). *Challenges in the no child left behind act for English language learners* (CRESST Policy Brief 7). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied linguistics, 14*(2), 115–129.

Blackledge, A. (2009). "As a country we do expect": The further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly, 6*(1), 6–16.

Bourdieu, P. (1991). *Language and symbolic power* (trans: Gino Raymond and Matthew Adamson). Cambridge, MA: Harvard University Press.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 147–170). Mahwah: Lawrence Erlbaum Associates.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3–18). Mahwah: Lawrence Earlbaum Associates.

Cheng, L., Watanabe, Y. & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah: Lawrence Erlbaum Associates. Code of Practice.

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104–122.

Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing, 28*(2), 221–249.

Collier, V., & Thomas, W. (2002). Reforming education policies for English learners means better schools for all. *The State Education Standard, 3*(1), 30–36.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing, 14*, 328–339.

De Jong, J., Lennig, M., Kerkhoff, A., & Poelmans, P. (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly, 6*(1), 41–60.

Eades, D. (2009). Testing the claims of asylum seekers: The role of language analysis. *Language Assessment Quarterly, 6*(1), 30–40.

Eribon, D. (1992). *Michel Foucault* (trans: Betsy Wing). Cambridge, MA: Harvard University Press.

Evans, B., & Hornberger, N. (2005). No child left behind: Repealing and unpeeling federal language education policy in the United States. *Language Policy, 4*(1), 87–106.

Extra, G., Spotti, M., & van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives*. London: Continuum.

Foucault, M. (1979). *Discipline and punish: The birth of the prison* (trans: from the French by Alan Sheridan). New York: Vintage Books.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational researcher, 18*(9), 27–32.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253–266.

Government Accountability Office. (2006). *No child left behind act: Assistance from education could help states better measure progress of students with limited English proficiency.* Washington, DC: Author.

Gysen, S., Kuijper, H., & van Avermaet, P. (2009). Language testing in the context of immigration and citizenship: The case of the Netherlands and Flanders. *Language Assessment Quarterly, 6*(1), 98–105.

Hanson, F. A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley: University of California Press.

Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *Contextualizing the age factor: Issues in early foreign language learning*. Berlin\New York: Mouton de Gruyter.

Kunnan, A. J. (2009). Testing for citizenship: The U.S. naturalization test. *Language Assessment Quarterly, 6*(1), 89–97.

Levi, T. (2015). Towards a framework for assessing foreign language oral proficiency in a large-scale test setting: Learning from DA mediation examinee verbalizations. *Language and Sociocultural Theory, 2*(1), 1–24.

Levi, T. (2016). Developing L2 oral language proficiency using concept-based Dynamic Assessment within a large-scale testing context. *Language and Sociocultural Theory, 3*(2), 197–220.

Levin, T., & Shohamy, E. (2008). Achievement of immigrant students in mathematics and academic Hebrew in Israeli school: A large-scale evaluation study. *Studies in Educational Evaluation, 34*(1), 1–14.

Levin, T., & Shohamy, E. (2012). Understanding language achievement of immigrants in schools: The role of multiple academic languages. In M. Leikin, M. Schwartz, & Y. Tobin (Eds.), *Current issues in bilingualism: Cognitive and socio-linguistic perspectives add page numbers* (pp. 137–155). Springer: Literacy Studies.

Levin, T., Shohamy, E., & Spolsky, B. (2003). *Academic achievements of immigrants in schools*, *Report submitted to the Ministry of Education (in Hebrew)*. Tel Aviv: Tel Aviv Univeresity.

Levin, T., Shohamy, E., & Inbar, O. (2007). Achievements in academic Hebrew among immigrant students in Israel. In N. Nevo & E. Olshtain (Eds.), *The Hebrew language in the era of globalization* (pp. 37–66). Jerusalem: Magnes Press, the Hebrew University.

Madaus, G. (1990, December 6). *Testing as a social technology*. Paper presented at the Inaugural Annual Boisi Lecture in Education and Public Policy, Boston College.

McNamara, T. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics, 18*, 304–319.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89–95.

Menken, K. (2006). Teaching to the test: How standardized testing promoted by *No Child Left Behind* impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal, 30*(2), 521–546.

Menken, K. (2007). High-stakes tests as de facto language policies in education. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education*, *Language testing and assessment* (Vol. 7, pp. 401–414). Netherlands: Kluwer.

Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Clevedon, Avon: Multilingual Matters.

Menken, K. (2010). *No Child Left Behind* and English language learners: The challenges and consequences of high-stakes testing. *Theory into Practice, 49*(2), 121–128.

Menken, K. (2013). Restrictive language education policies and emergent bilingual youth: A perfect storm with imperfect outcomes. *Theory into Practice, 52*(3), 160–168.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*, 9–20.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 447–474). New York: ACE/Macmillan.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series, 1996* (1), i–18.

Milani, T. M. (2008). Language testing and citizenship: A language ideological debate in Sweden. *Language in Society, 37*(01), 27–59.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173.
Schissel, J. (2012). *The pedagogical practice of test accommodations with emergent bilinguals: Policy-enforced washback in two urban schools* (Unpublished doctoral dissertation). Philadelphia: University of Pennsylvania.
Schupbach, D. (2009). Testing language, testing ethnicity? Policies and practices surrounding the ethnic German *Aussiedler*. *Language Assessment Quarterly, 6*(1), 78–82.
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*, 340–349.
Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*, 331–345.
Shohamy, E. (2001a). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education.
Shohamy, E. (2001b). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391.
Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. Abington: Oxon Abingdoi.
Shohamy, E. (2009). Language tests for immigrants: Why language? Why tests? Why citizenship? In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration* (pp. 45–59). Amsterdam: John Benjamins.
Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal, 95*(3), 418–429.
Shohamy, E. (2015). *Critical language testing and English Lingua Franca: How can one help the other? Waseda Working Papers in ELF (English as a Lingua Franca)* (Vol. 4, pp. 37–51). Waseda ELF Research Group Waseda University: Tokyo.
Shohamy, E., & Kanza, T. (2009). Citizenship, language, and nationality in Israel. In G. Extra, M. Spotti, & P. van Avermaet (Eds.), *Language testing, migration and citizenship: Cross-national perspectives*. London\New York: Continuum.
Shohamy, E., & McNamara, T. (2009). Language tests for citizenship, immigration and asylum. *Language Assessment Quarterly, 6*(1), 1–5.
Shohamy, E., & Menken, K. (2015). Language assessment: Past and present misuses and future possibilities. Bi-multi-lingual assessment. The Routledge handbook on bilingual education. In W. E. Wright, S. Boun, & O. García (Eds.), *The handbook of bilingual and multilingual education* (1st ed.). London/New York: Routledge.
Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317.
Solano-Flores, G., & Trumball, E. (2003). Examining language in context: The need for new research paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13.
Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.
Stevenson, P. (Ed.) (2009). 'National' languages in transnational contexts: Language, migration and citizenship in Europe. In: *Language ideologies, policies and practices: Language and the future of Europe* (pp. 147–161). London: Palgrave Macmillan.
Valdes, G., & Figueroa, R. (1996). *Bilingualism and testing: A special case of bias*. Norwood: Ablex.
Valdés, G., Menken, K., & Castro, M. (2015). *Common Core, bilingual and English language learners: A resource for educators*. Philadelphia: Caslon Publishing.
Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41–69.

# Index



© Springer International Publishing AG 2017

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1