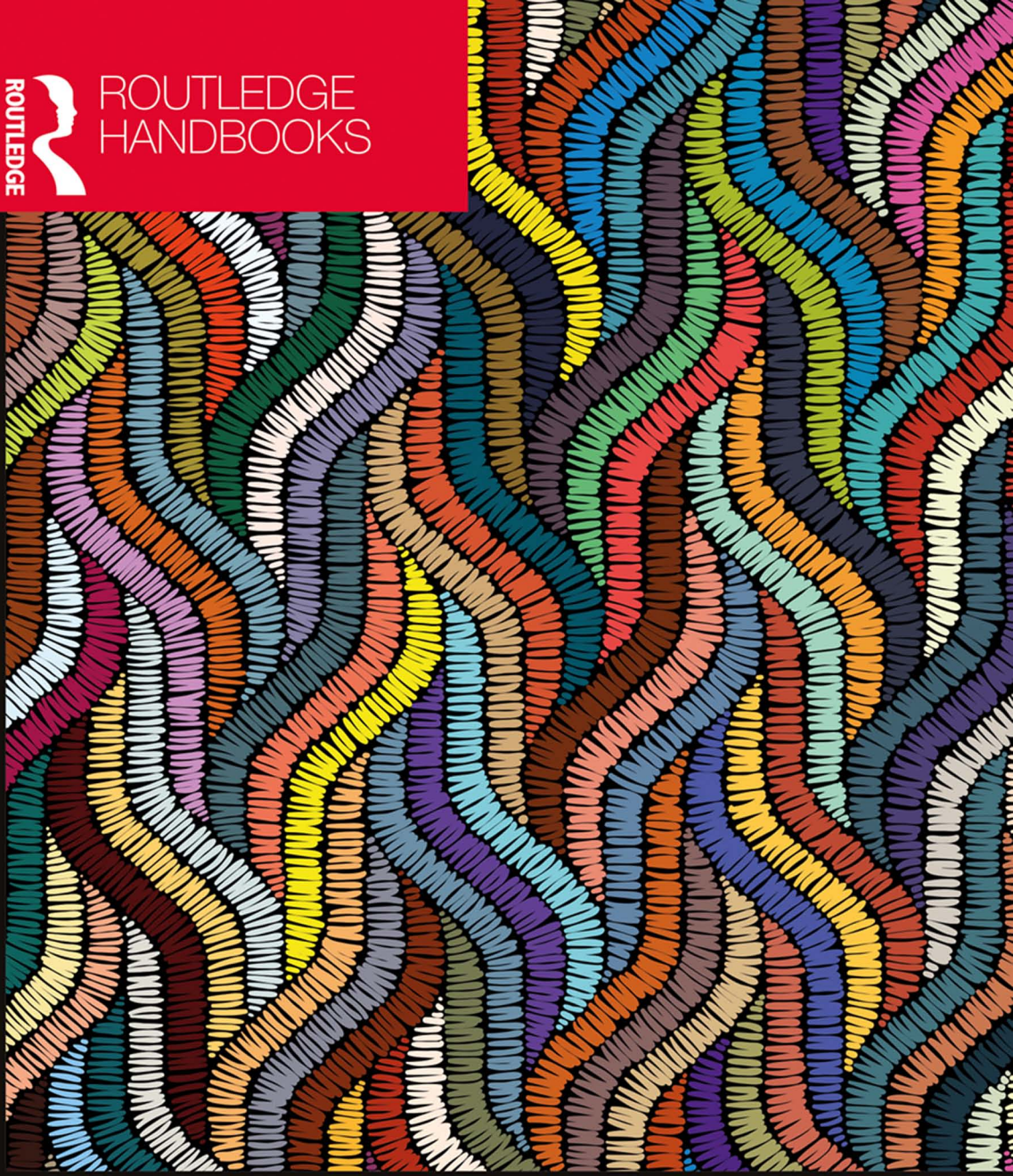
ROUTLEDGE HANDBOOKS



The Routledge Handbook of Language Testing

Second Edition

Edited by Glenn Fulcher and Luke Harding

The Routledge Handbook of Language Testing

This second edition of *The Routledge Handbook of Language Testing* provides an updated and comprehensive account of the area of language testing and assessment.

The volume brings together 35 authoritative articles, divided into ten sections, written by 51 leading specialists from around the world. There are five entirely new chapters covering the four skills: reading, writing, listening, and speaking, as well as a new entry on corpus linguistics and language testing. The remaining 30 chapters have been revised, often extensively, or entirely rewritten with new authorship teams at the helm, reflecting new generations of expertise in the field. With a dedicated section on technology in language testing, reflecting current trends in the field, the *Handbook* also includes an extended epilogue written by Harding and Fulcher, contemplating what has changed between the first and second editions and charting a trajectory for the field of language testing and assessment.

Providing a basis for discussion, project work, and the design of both language tests themselves and related validation research, this *Handbook* represents an invaluable resource for students, researchers, and practitioners working in language testing and assessment and the wider field of language education.

Glenn Fulcher is Professor of Applied Linguistics and Language Assessment at the University of Leicester (UK). He has served as president of the International Language Testing Association and as editor of the journal *Language Testing*. His Routledge book *Re-examining Language Testing* was joint winner of the SAGE/ILTA book award, together with the first edition of this *Handbook*.

Luke Harding is Professor in Linguistics and English Language at Lancaster University (UK). His research interests are in applied linguistics and language assessment, particularly assessing listening and speaking, World Englishes and English as a Lingua Franca, and language assessment literacy and professional ethics. He is currently co-editor of the journal *Language Testing*.

Routledge Handbooks in Applied Linguistics

Routledge Handbooks in Applied Linguistics provide comprehensive overviews of the key topics in applied linguistics. All entries for the handbooks are specially commissioned and written by leading scholars in the field. Clear, accessible and carefully edited *Routledge Handbooks in Applied Linguistics* are the ideal resource for both advanced undergraduates and postgraduate students.

THE ROUTLEDGE HANDBOOK OF CORPUS APPROACHES TO DISCOURSE ANALYSIS
Edited by Eric Friginal and Jack A. Hardy

THE ROUTLEDGE HANDBOOK OF WORLD ENGLISHES
Second Edition
Edited by Andy Kirkpatrick

THE ROUTLEDGE HANDBOOK OF LANGUAGE, GENDER AND SEXUALITY
Edited by Jo Angouri and Judith Baxter

THE ROUTLEDGE HANDBOOK OF PLURILINGUAL LANGUAGE EDUCATION
Edited by Enrica Piccardo, Aline Germain-Rutherford and Geoff Lawrence

THE ROUTLEDGE HANDBOOK OF THE PSYCHOLOGY OF LANGUAGE LEARNING AND TEACHING
Edited by Tammy Gregersen and Sarah Mercer

THE ROUTLEDGE HANDBOOK OF LANGUAGE TESTING
Second Edition
Edited by Glenn Fulcher and Luke Harding

For a full list of titles in this series, please visit www.routledge.com/series/RHAL

The Routledge Handbook of Language Testing

Second Edition

Edited by Glenn Fulcher and Luke Harding

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Second edition published 2022
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge 2012

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Fulcher, Glenn, editor. | Harding, Luke, editor.
Title: The Routledge handbook of language testing / edited by Glenn Fulcher and Luke Harding.
Description: Second edition. | London ; New York : Routledge, 2021. | Series: Routledge handbooks in applied linguistics | Includes bibliographical references and index.
Identifiers: LCCN 2021021439 | ISBN 9781138385436 (hardback) | ISBN 9781032116501 (paperback) | ISBN 9781003220756 (ebook)
Subjects: LCSH: Language and languages—Ability testing. | LCGFT: Essays.
Classification: LCC P53.4 .R68 2021 | DDC 401/.93—dc23
LC record available at https://lccn.loc.gov/2021021439

ISBN: 978-1-138-38543-6 (hbk)
ISBN: 978-1-032-11650-1 (pbk)
ISBN: 978-1-003-22075-6 (ebk)

DOI: 10.4324/9781003220756

Typeset in Bembo
by Apex CoVantage, LLC

Contents

Figures

Tables

Contributors

Nwara Abdulhamid, PhD, completed her doctoral study of high-stakes testing, curricular alignment, and washback on teaching and learning in Applied Linguistics at Carleton University. She currently engages in research regarding language education, teacher education and educational assessment, and relationships between misalignment and washback in high-stakes testing contexts.

Jamal Abedi is a professor at the School of Education of the University of California, Davis, and a research partner at the National Center for Research on Evaluation, Standards, and Student Testing. His research interests include accommodations and classification for English language learners and comparability of alternate assessments for students with significant cognitive disabilities.

Marta Antón is Professor of Spanish at Indiana University-Purdue University Indianapolis. Her main research focus is the application of sociocultural theory to second language learning and teaching in the classroom, including teacher–learner interaction, peer interaction, and dynamic assessment. Other interests are classroom-based assessment, authentic assessment, and Spanish sociolinguistics.

James Dean Brown is Professor Emeritus in the Department of Second Language Studies at the University of Hawai'i at Mānoa. He has spoken and taught courses in places ranging from Albuquerque to Zagreb and has published numerous articles and books on language testing, curriculum design, research methods, and connected speech.

Tineke Brunfaut is Professor in Linguistics and English Language at Lancaster University, UK. Her main research interests are language testing and reading and listening in a second/foreign language. In recent research, she has focused on cognitive and affective factors and processes in language testing, methodology in language testing research, and language test development.

Yuko Goto Butler is Professor of Educational Linguistics in the Graduate School of Education at the University of Pennsylvania. She is also the director of the Teaching English to Speakers of Other Languages (TESOL) program at Penn. Her research interests include language assessment and second and bilingual language learning among children.

Carol A. Chapelle is Distinguished Professor in Liberal Arts and Sciences at Iowa State University. Recent books include *Validity Argument in Language Testing: Case Studies of*

Validation Research (Cambridge, 2021; with E. Voss), *Argument-Based Validation in Testing and Assessment* (Sage, 2021), and *The Handbook of Technology and Language Learning and Teaching* (Wiley, 2017; with S. Sauro).

Liying Cheng is Professor and Director of the Assessment and Evaluation Group (AEG) at the Faculty of Education, Queen's University. Her primary research interests are the impact of large-scale testing on instruction, the relationships between assessment and instruction, and the academic and professional acculturation of international and new immigrant students, workers, and professionals. Her seminal research on washback focuses on the global impact of large-scale testing.

Andrew D. Cohen, Professor Emeritus, University of Minnesota, has conducted research for many years on language learner strategies, including a focus on how learners respond to tests of reading (cloze and multiple choice) and oral language skills (e.g., skills in pragmatics). He uses verbal report techniques to explore test response strategies.

Sara T. Cushing is Professor of Applied Linguistics at Georgia State University. She has published research in the areas of language assessment, second language writing, and teacher education. She has been invited to speak and conduct workshops on second language assessment throughout the world, most recently in Vietnam, Colombia, and Thailand.

Larry Davis is a research scientist in the Center for Language Education, Assessment, and Research (CLEAR) in the research and development division of Educational Testing Service. His interests broadly encompass the assessment of speaking, including creation of rubrics; rater cognition; automated evaluation of speech; and assessment of spoken interaction.

Bart Deygers is an assistant professor of second language didactics and assessment at Ghent University, Belgium. In recent research, he has focused on language testing policy for migration purposes, the use of the CEFR in migration policy, and the impact of language requirements on low-literate migrants.

Kathrin Eberharter holds a research position at the University of Innsbruck. Her main interests are in the area of assessing L2 writing and speaking, with a focus on task development, rating scale development, rater cognition, and training. She is also interested in cognitive processes in language testing and language assessment literacy.

Jason Fan is the deputy director and senior research fellow at the Language Testing Research Centre at the University of Melbourne. His research interests include test validity and validation, research methods, and language assessment literacy.

Janna Fox is Professor Emerita of Applied Linguistics and Discourse Studies in the School of Linguistics and Language Studies, Carleton University. Her research interests include language assessment, within and external to the classroom; washback and impact on teaching and learning; and sociocultural, transdisciplinary perspectives in validation research.

Glenn Fulcher is Professor of Applied Linguistics and Language Assessment at the University of Leicester (UK). He has served as president of the International Language

Testing Association and as editor of the journal *Language Testing*. His Routledge book *Re-examining Language Testing* was joint winner of the SAGE/ILTA book award together with the first edition of this *Handbook*.

Evelina D. Galaczi is Head of Research Strategy at Cambridge Assessment English and has worked in language education for over 30 years as a researcher, assessment specialist, teacher, and materials writer. Evelina's expertise lies in speaking assessment, interactional competence, test development, and the use of technologies in assessing productive skills.

Próspero N. García is associate professor of Spanish applied linguistics at Rutgers University, Camden. His research focuses on sociocultural psychology applied to heritage and second language acquisition and pedagogy, language evaluation and assessment, teacher's cognition, and technology-enhanced language learning.

Gene B. Halleck retired as Professor of Linguistics/TESL at Oklahoma State University. She has designed a set of diagnostic tests, the Oral Proficiency Tests for Aviation, to accompany an aviation English curriculum designed for the US Federal Aviation Administration.

Luke Harding is Professor in Linguistics and English Language at Lancaster University (UK). His research interests are in applied linguistics and language assessment, particularly assessing speaking and listening, World Englishes and English as a Lingua Franca, and language assessment literacy and professional ethics. He is currently co-editor of the journal *Language Testing*.

Franz Holzknecht is a researcher in applied linguistics and language testing at the University of Applied Sciences in Special Needs Education, Zurich, Switzerland. His main research interests are in language testing, particularly in the areas of sign language testing; testing second language listening; and language test development.

Chihiro Inoue is Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her main research interests are in designing tasks and rating scales for L2 speaking and the effects of task-related variables on learner language.

Eunice Eunhee Jang is Professor at Ontario Institute for Studies in Education, University of Toronto. Her research areas include diagnostic assessment for supporting students and teachers, validation of English proficiency descriptors-based assessment for K–12 English language learners and scenario-based language assessment, evaluation of school effectiveness in challenging circumstances, and technology-infused assessment system design and validation.

Yan Jin is Professor of Applied Linguistics of the School of Foreign Languages at Shanghai Jiao Tong University, China. Her research interests are in the development and validation of large-scale and high-stakes language assessments. She has been chair of the National College English Testing Committee since 2004.

Michael T. Kane is the Messick Chair in validity at the Educational Testing Service. His research interests are validity theory, generalizability theory, and standard setting. Dr. Kane

holds a PhD in education, an MS in statistics from Stanford University, and a BS in physics from Manhattan College.

Nahal Khabbazbashi is Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her main research interests are the assessment of speaking, the use and impact of technology in assessment, and multimodal communication.

Ute Knoch is Associate Professor in Language Assessment and the Director of the Language Testing Research Centre at the University of Melbourne. Her research interests are in the areas of writing assessment, rating processes, rating scale development, test validation, assessing languages for academic and professional purposes, and placement testing.

Benjamin Kremmel is head of the Language Testing Research Group Innsbruck (LTRGI) at the University of Innsbruck, Austria, where he teaches and researches language learning, teaching, and assessment. His research interests include vocabulary assessment, language assessment literacy, diagnostic language testing, and the interface between language testing and SLA.

Antony John Kunnan (PhD, UCLA) is a principal assessment scientist at Duolingo, Inc. He specializes in the areas of fairness, validation, and ethics. He is the author and editor of over 90 publications; he recently authored a book titled *Evaluating Language Assessments* (Routledge, 2018). He is a past president of the International Language Testing Association and the founding president of the Asian Association for Language Assessment. He is also the founding editor of *Language Assessment Quarterly*.

Hye-won Lee is Senior Research Manager at Cambridge Assessment English. She holds a PhD in Applied Linguistics and Technology from Iowa State University, with specialization in technology-enhanced language assessment and argument-based validation. Her current work focuses on assessing speaking via video call and defining language ability in data-driven diagnostic assessment.

Gad S. Lim is Director of Assessment at Michigan Language Assessment. His areas of interest include the design of learning-oriented assessment, the development and setting of standards, and the application of these in education policy and reform.

Sarah McElwee is Principal Research Manager in the Research and Thought Leadership group at Cambridge Assessment English, University of Cambridge. Her research interests include consequential validity; cognitive factors in testing, particularly related to younger learners; and the role of validation in product lifecycle management for assessment.

Robert J. Mislevy is the Frederic M. Lord Chair in Measurement and Statistics at ETS and Professor Emeritus of Measurement, Statistics, and Evaluation at the University of Maryland, with affiliation in second language acquisition. His research applies developments in statistics, technology, and cognitive science to practical problems in educational assessment.

Carol Lynn Moder is Professor of Linguistics and TESOL at Oklahoma State University. She has designed ESP testing and curriculum materials for International Teaching Assistants,

English for Academic Purposes, and aviation English. She designed aviation English assessments for the international training division of the US Federal Aviation Administration and for Ordinate Corporation. She served as a consultant to ICAO in 2005.

Fumiyo Nakatsuhara is Reader in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her main research interests include the nature of co-constructed interaction in speaking tests, task design, rating scale development, and the relationship between listening and speaking skills.

Susan Nissan is Senior Director at Educational Testing Service (retired) in Princeton, New Jersey. In addition to assessment development, her research interests include test design, test validity, standard setting, assessment literacy, and the relationships between assessment and learning.

Gary J. Ockey is a professor of Applied Linguistics and Technology at Iowa State University. He is interested in the use of technology for aiding in the assessment of second language oral communication and the applications of various quantitative statistical approaches to facilitate the understanding and practice of second language education and assessment.

Elizabeth Park is an assessment specialist at Educational Testing Service. She designs English language proficiency assessments and professional development programs for teachers of English as a second/foreign language.

Lia Plakans is a professor of Multilingual Education at the University of Iowa in the US. Her research interests include integrated skills assessment, assessing reading and writing, test use, and bilingual assessments in school settings.

John Read is an emeritus professor in Applied Language Studies at the University of Auckland. His research interests include the design of second language vocabulary tests, academic literacy and the post-admission assessment of university students, and testing English for occupational purposes, particularly in aviation and the health sciences.

Steven J. Ross is a professor of second language acquisition at the University of Maryland. His recent research interests are assessment of second language pragmatics, validity theory, language assessment, and longitudinal research methods. Professor Ross completed a PhD in second language acquisition at the University of Hawai'i Manoa in 1995.

Nick Saville is Director of Research and Thought Leadership (Cambridge Assessment English, University of Cambridge) and Secretary-General of the Association of Language Testers in Europe. His research interests include assessment and learning in the digital age, the use of ethical AI, language policy and multilingualism, the CEFR, and learning-oriented assessment.

Yasuyo Sawaki is Professor of Applied Linguistics at Waseda University in Tokyo, Japan. She has research interests in diverse topics in second/foreign language assessment, including test validation, assessment and instruction of summary writing, diagnosing language ability, and using technology in language assessment.

Dongil Shin is Professor of Applied Linguistics in the Department of English Language and Literature, Chung-Ang University, Korea. Since completing his doctorate at University of Illinois at Urbana-Champaign, he has researched language testing (policies), (critical) discourse analysis, and language ideologies and subjectivities, mostly in Korean contexts. He tries to explore critical discursive approaches to language testing in newly emerging multilingual (Korean) societies.

Jeanne Sinclair earned her PhD in 2020 from the Ontario Institute for Studies in Education (OISE) of the University of Toronto. She currently serves as senior psychometrician with the Nursing Community Assessment Service (British Columbia) and as postdoctoral fellow (OISE/McMaster University) researching child development and educational stratification.

Nasreen Sultana has earned PhD in Education from Queen's University, Kingston, with a dissertation exploring the relationship between washback and curriculum alignment. Currently, she is a full-time teaching and learning consultant at Conestoga College, Kitchener, Ontario. Nasreen's areas of interest include washback, intercultural pedagogy, equity, diversity, and inclusion.

Carolyn E. Turner is Associate Professor in the Department of Integrated Studies in Education, McGill University (retired). Her research interests include language assessment in educational and multi-lingual health professional settings, the impact of high-stakes tests on teaching and learning, learning-oriented assessment, empirically based rating scales, and mixed-methods research designs.

Elvis Wagner is Associate Professor of TESOL at Temple University, where he coordinates the PhD in Applied Linguistics program and the World Languages Education program. His research interests include the assessment of L2 oral ability, specifically focusing on how L2 listeners process and comprehend unscripted, spontaneous spoken language.

F. Scott Walters is Associate Professor in the Office of International Education at Seoul National University of Science and Technology. His research interests include the development and validation of L2 pragmatics tests, conversation analysis, and assessment-literacy development.

Xiaoming Xi is Chief of Product, Assessment, and Learning, VIPKID International. Her areas of interest include validity, validation frameworks, fairness frameworks, test design, speaking assessment, human scoring, rater issues, automated scoring and feedback, and the use of AI technology in learning and assessment.

Xun Yan is an associate professor of Linguistics, Second Language Acquisition and Teacher Education (SLATE), and Educational Psychology at the University of Illinois at Urbana-Champaign. His research interests include speaking and writing assessment, psycholinguistic approaches to language testing, and language assessment literacy.

Chengbin Yin specializes in Second and Foreign Language Education and Measurement, Statistics, and Evaluation at the University of Maryland, College Park. Her research interests are assessment design from a socio-cognitive perspective and measurement models in educational assessment.

Richard F. Young is Professor Emeritus of English and Second Language Acquisition at the University of Wisconsin-Madison. His research is concerned with the relationship between language and social context, with a particular focus on variation in interlanguage, morphology, talking and testing, language and interaction, and discursive practice in language teaching and learning.

Acknowledgements

We would first like to pay tribute to Fred Davidson, one half of the original editorial team. Fred's work on the first edition laid extremely strong foundations for the second. It was an honour to carry on this work.

Second, we thank our brilliant contributors. There were 51 authors involved in this second edition (not including us), and everyone has been a pleasure to work with. The atmosphere around this project has always been convivial and collaborative, and we thank all contributors for making our task easy.

We would also like to thank Hannah Rowe and Eleni Steck at Routledge, who have assisted at various stages of this project, Kate Fornadel at Apex CoVantage, and Geisa Davila Perez (Lancaster University) for some last-minute editorial assistance.

Finally, on a more personal note, Glenn thanks Jenny for sharing all the lovely things in life. And Luke thanks Rachael, Iris, and Saffron for being great company during the various lockdowns of 2020–2021, for their constant support, and for always making life fun.

Editorial

The first edition of *The Routledge Handbook of Language Testing* was published in 2012. In 2016, it was the joint winner of the SAGE/International Language Testing Association prize for the best book in language testing. The international selection committee provided the following citation to justify the award:

> "[T]he editors have succeeded in assembling a set of contributors with an unparalleled level of expertise in their respective areas, and with distinctive talents in communication. The strength of this extremely well-edited collection lies in the interweaving of theoretical and practical aspects of language testing through nine broad themes, and in the structuring of individual contributions to provide a historical perspective, a discussion of current issues and contributions, and a consideration of future directions. The volume stands not only to have a wide impact on best practice in the field, but also in the development of language assessment literacy in other professionals who find themselves involved in activities of language assessment."

Adoption by teachers and learners alike has seen the *Handbook* become the standard reference text in the field, but theory and research do not stand still, and in the last decade, the field has continued to flourish and expand with the use of new technologies, assessment purposes, and contexts of score use. The rapid expansion of research in language assessment literacy has also provided information about what is required in reference works that underpin successful pedagogy for language teachers, future language testers, and other stakeholders. In preparing for this second edition, we therefore not only took into account changes in research and assessment literacy needs, but also engaged with the publishers in a survey of users to discover what changes they would like to see. Perhaps the largest change is the introduction of Section 4 on assessing the language skills, as skills assessment remains a key aspect of many language testing programmes around the world, but there are numerous modifications to both content and focus throughout the volume.

DOI: 10.4324/9781003220756-1

We also welcome many new authors, reflecting the ever-expanding and international nature of language testing research. Some of the contributors to the first volume have retired, and we have sadly seen the passing of Alan Davies, a towering academic thinker who has left an outstanding legacy to the language testing field and literature. In the words of Marcus Aurelius, "Flows and changes are constantly renewing the world, just as the ceaseless passage of time makes eternity ever young" (Meditations 4, 36), and so it is with research and our understanding of assessment theory and practice. We have therefore strived to include a diverse and exciting authorship that retains the insight of established scholars alongside the novelty of recent innovations and insights from the next generation of researchers.

What has not changed is the editorial philosophy that made the first edition such a great success. We have always believed that the role of editors is to provide structure and direction to the volume and to aid authors in executing their arguments coherently and cohesively. The editor's role is not to direct content, arguments, or conclusions, although they may make recommendations that help strengthen chapter coverage. All too often, editors wish to shape publications in their own image, but this is not how fields develop or individuals learn. As J. S. Mill would argue, attempts to control reveal a presumption of infallibility, and for true progress, all facts and interpretations must be heard and discussed with open critical minds. Only in this manner can we ultimately provide warrants to justify an emerging consensus that at some future point will again be challenged, and so on endlessly (Mill, 1859: 25–26). Such is the nature of learning and progress.

This is, of course, the rationale for future editions of this *Handbook*, but we will refrain from crystal-ball gazing in the editorial and leave that to the predictions offered by our authors and to our Epilogue to this volume. We therefore turn to the updated content, which we believe provides the same authoritative, stimulating, and pedagogically useful platform as the first edition for investigating the field of language testing and assessment.

Content

Section 1 Validity

In Chapter 1, Chapelle and Lee provide an evolutionary account of validity theory and the practice of validation. As they make clear, there are many disparate ways of understanding validity. The one that researchers choose is either quite random, depending on their background and reading, or very deliberate, based on a commitment to an underpinning philosophy. Chapelle and Lee are unequivocal in adopting an argument-based approach as the lens through which to view the evolution of theory and practice, which carries with it an instrumental concern with the day-to-day practice of conducting validity research. They argue that it provides both a common language and the specific tools that the field needs to move forward. It is unquestionably the case that at the present time, the argument-based approach to validity and validation is in the ascendency, and it is therefore right and proper that this chapter appears first in the second edition of the *Handbook*. But readers should be aware that it is but one option in the current marketplace of ideas (Fulcher, 2015: 108–12).

Chapter 2 is the most natural progression from the position presented in Chapter 1, representing as it does the clearest possible statement of an argument-based approach to validation by its principal proponent, Michael T. Kane. He sets out the two fundamental building blocks of validation. First comes the interpretative/use argument (IUA), which establishes the proposed uses of test scores and the interpretations of those scores required for their

use. Second is the validity argument that evaluates the plausibility of the IUA by evaluating the evidence and rationales provided to support score use. The chapter explains how a test developer would go about building a confirmatory research agenda and a critical evaluator conduct studies that investigate alternative plausible interpretations of score meaning. The reader should note that Kane focuses on *observable attributes*, which are contrasted with *traits*. The latter are underlying theoretical abilities that account for observed behaviour, but in argument-based approaches, the preference is to focus on what can be observed and how consistent behaviours associated with target domains can be rendered into a score. Kane illustrates his approach with examples that will be familiar to language teachers and testers, making this chapter a highly accessible introduction to mainstream argument-based validation.

In Chapter 3, Ross begins his discussion with an acknowledgement that the argument-based approach to validation has done a great deal to help structure validation evidence in such a way that it can be evaluated in relation to counter-explanations. This is an important observation, as it pinpoints the main contribution made by the appropriation of Toulmin's argument structure to language test validation. But drawing inferences from evidence is often problematic, and Ross carefully analyses the pitfalls that validation researchers face with respect to both quantitative and qualitative data. Anyone who still believes that the outcomes of statistical analysis represent a universally generalisable truth should read and reread this chapter. Its treatment of how we make inferences from scores and, more generally, in research is one of the most masterful in the language testing literature, alongside Bachman (2006). Readers of the first edition of the *Handbook* will note that we have moved this revised chapter into the section on validity because of the major contribution this revised chapter makes to validation theory and practice.

Section 2 The uses of language testing

When we invited Richard F. Young to revise his chapter for the new edition of the *Handbook*, he declined – quite understandably – on account of now being retired. After some consideration, we decided to include the original chapter again (with Richard's consent, of course), with a few minor alterations. Following Cathie Elder's review of this chapter, we have also moved it into the first position within this section, as Chapter 4. There has been plenty of research on the consequences of language tests in the years since the first edition; much of this research is picked up in other chapters in the volume written by Deygers, Walters, and Cheng and Sultana, among others. Similarly, there have been important contributions to our understanding of interactional competence in language contexts. Again, this research is discussed in a new chapter by Nakatsuhara, Khabbazbashi, and Inoue. The reason that we decided to include this chapter again is that we believe it is utterly unique. Young's discussion is wide ranging and, given that it was written more than a decade ago, eerily prescient. Consider, for example, his final paragraph, in which he imagines what the future of language testing might look like: "an image of two psychometricians, experts in the field of educational measurement, sitting in front of a computer monitor scratching their heads as a waterfall of data pours down the screen." But we include this chapter, not because it is a historical artefact (after all, the interested reader could go back to the 2012 edition and read it there). Rather, we include it because it makes sense within this current collection and draws together various themes that are still highly relevant for language testing and assessment. Leaving this chapter out would have diminished the volume. If you haven't read it yet, we strongly encourage you to do so now.

It is arguably the case that language testing for specific purposes is the most high-stakes assessment that we find. In Chapter 5, Moder and Halleck consider the history and practice of aviation English language testing. Perhaps two critical issues from the many involved come to the fore. The first is the interaction, negotiation, and inevitable compromise between language testers, institutions that make policy, and very powerful stakeholders. The second is the requirement that domain-specific language tests be firmly grounded in an analysis of the language used in the workplace context as defined in the purpose of the test. This is the cutting edge of interaction between applied linguistic investigation and language assessment. Moder and Halleck demonstrate the critical role that language testing plays in maintaining safety for the public in very high-stakes social contexts, and their discussion generalises to many other situations in which the risk associated with error is particularly grave. During the COVID-19 pandemic, this language of error, risk, and consequence has become the staple of news items, and their consideration is no less serious in language assessment than it is in the field of medicine.

There is no topic in language testing more controversial than the use of tests for making decisions about immigration and citizenship. Despite many well-reasoned studies critiquing the use of language tests for these purposes – from both within and outside the field of language testing – the practice continues and appears to be proliferating across numerous countries. Is there a fair and ethical way to conduct language assessment for immigration and citizenship purposes? In Chapter 6, Kunnan revisits his contribution to the first edition of the *Handbook* by considering one particular gatekeeping exam in detail: the United States Naturalization Test. Kunnan provides a vivid description of the historical antecedents to the introduction of the test before scrutinising the current version, evaluating it against principles of fairness and justice. The test does not emerge well from this critique, and Kunnan's chapter demonstrates how a structured and systematic approach to fairness and justice can provide just the right tools for highlighting a flawed and potentially harmful testing regime.

Section 3 Classroom assessment and washback

For a long time, there has been an imbalance in language testing research, with a strong focus on large-scale international tests at the expense of a much more common activity: classroom-based language assessment. In classrooms around the world, teachers regularly set assessments, mark tests, provide feedback, and prepare learners for high-stakes exams. And yet there is so much we don't know or understand about assessment practices in these everyday contexts. In Chapter 7, Fox and Abdulhamid join Turner (who was the sole author of the chapter in the first edition) to provide an up-to-date overview of classroom-based assessment, charting what has changed over the past decade and what issues remain the same. This chapter illustrates that classroom-based language assessment has come a long way quickly, with newer unifying frameworks such as learning-oriented assessment helping drive a classroom-based assessment "turn." At the same time, technology is having an important impact on classroom assessment as digital tools provide some relief for teachers (e.g., automated writing evaluators), as well as a range of new challenges. The general paradigm shift toward a greater blending of external, large-scale testing and teacher assessment is a key feature of this chapter. In another ten years' time, we might expect to see technology mediating a much smoother alignment between classroom assessment and standardised assessment, though it will be interesting to observe how this is done and whose needs and objectives are prioritised: those of learners, teachers, policy makers, or test providers?

Washback is an area of language testing that has received sustained attention since Alderson and Wall's landmark paper in 1993, "Does washback exist?" Many of the hypotheses raised in that article have since been investigated empirically, and the findings have clarified our understanding. A fair summary would seem to be yes, washback exists, but it's complicated. Cheng and Sultana explore some of this complexity in Chapter 8, focusing on literature which has been published in the period since the first edition of the *Handbook*. The authors focus on three main trends in washback research over the past decade: the expansion of washback studies into previously under-researched educational contexts, the connection of washback in the classroom with broader social/educational contexts, and the adoption of a wider range of conceptual/theoretical frameworks to conduct washback research. A key theme that emerges in this chapter is the importance of "alignment": that is, the extent to which teaching, curriculum, and assessment are in agreement. An increasing number of studies now show that changing a test is not enough, on its own, to bring about positive washback. In fact, *negative* washback is the more likely result in contexts in which there is no alignment between the test, the curriculum goals, and classroom practices. We hope that policy makers in language education contexts will read this chapter and note that the introduction of a test, on its own, rarely solves problems and quite often creates new ones.

Since the first edition of the *Handbook*, assessment of young learners has expanded rapidly. We see this in the range of commercial tests now on the market targeted at children and teenagers, and we also see it in the number of research articles and new books focusing on young learners. The increased understanding of how best to assess young learners is a very welcome development, particularly given that learners under 18 are likely to comprise a large proportion of the number of language learners worldwide. In Chapter 9, Butler brings great expertise to this topic and provides a thorough and comprehensive overview of the current state of play in assessing this test-taker population. Butler highlights the challenges of dealing with great degrees of variability within the characteristics of young learners and also points toward the rapid shift of younger generations toward digital technologies. Previous reservations about test taker "computer familiarity" seem long gone in many contexts; in their place are concerns that test developers may not be able to keep up with the digital communicative practices of learners in this age range. As other authors in this volume note, however, technology in language testing also has its limitations. At the end of the chapter, Butler cautions that technology cannot replace teachers and emphasises the importance of teachers' "diagnostic competence" in young learner assessment. Young learner assessment looks set to remain a site of conceptual and practical advances in the coming years; we will watch these developments with interest.

Dynamic assessment has been a "hot" topic in language assessment for some time now. It offers a very radical departure from traditional testing practice, focusing on learning potential and the key role played by mediation. This places it at odds with more psychometric views which see language ability as residing solely within the mind of the individual learner. However, the dynamic assessment approach embodies an important critique of these more traditional testing practices. Scholarship in dynamic assessment forces us to consider that, if language ability develops through social interaction, then such interaction should play a primary role in our assessment practices. In Chapter 10, Antón and García update Antón's first-edition version, integrating research findings that have propelled dynamic assessment in new directions. Computerized dynamic assessment (CDA) features strongly; the authors describe fascinating new research in which new technology (such as messaging apps) is harnessed for its interaction and mediation potential. In the years to come, we expect to see the principles of dynamic assessment employed in many more assessment settings.

Diagnostic assessment is a feature of many professional contexts, most notably healthcare, but also IT support, car mechanics, engineering, and so on. Language testing has grappled with its own approach to diagnostic assessment for some time, but in recent years, there has been considerable progress on the topic through work on theory building and on technical issues (particularly in the area of cognitive diagnostic modeling – CDM). In Chapter 11 – an update of Jang's sole-authored contribution to the 2012 edition – current issues in diagnostic assessment are surveyed, and a range of novel directions are discussed. Jang and Sinclair provide an exciting glimpse into what diagnostic assessment might look like in the future, one in which machine learning could play an important role in capturing and processing data. Still, even with the most sophisticated methods for tracking and measuring learning, diagnostic assessment is not truly *diagnostic* without well-developed feedback systems. Here, Jang and Sinclair remind us that feedback utilization is a core component of effective diagnostic assessment and that evidence of use of feedback is essential for supporting validity arguments for diagnostic assessment. Our field needs more research on the effectiveness of specific, innovative diagnostic procedures. We hope that future researchers will read Jang and Sinclair's article and be inspired by the potential that diagnostic assessment offers.

Section 4 Assessing language skills

In the introduction to Chapter 12, Nakatsuhara, Khabbazbashi, and Inoue wisely quote Lado's observation that the ability to speak in a second language is the most prized objective of language learning. This is as true today as it has always been. Yet the assessment of speaking also remains as problematic as it was in the 1960s, despite all that we have learned since then, and is expertly documented in the historical section of this chapter. The authors consider the new range of task types available to us and the constructs of interest that each is claimed to reveal for scoring. As the scoring mechanism embodies the construct, it is not surprising that a variety of approaches have been proposed. The variety of competing claims about approaches also makes assessing speaking a controversial area of research and practice. In looking to the future, the authors state that technology has played a large role in shaping the speaking construct and will likely continue to do so. We also wonder whether this will be driven by commercial need, as it was during the COVID-19 pandemic, and how the machine scoring of speaking might model our humanity in the coming decades.

In another new addition to the *Handbook*, Chapter 13, Wagner provides a comprehensive overview of current challenges and debates in second language listening assessment. It is well understood among practitioners that listening is a complex skill to assess, both in terms of the practicalities of developing listening tests (sourcing or creating recordings, administering tests in contexts where resources are scarce) and in defining and operationalising the construct itself. But there is now a lot of research on listening assessment, and we were pleased to see Wagner "bust" the myth that listening is an under-researched skill. In fact, research on listening has grown over the past two decades, and in specific areas like the use of video in listening assessment and the role of speaker accent, there are now established canons of research to guide test developers in making good decisions about their own tests. The key problem is transforming that research into practice: designing innovative, authentic tasks that move beyond the traditional scripted, monologic, multiple-choice format. Wagner covers these issues, among many others, concluding, in a similar way to the previous chapter, that technology is likely to have a transformative approach on listening in the future.

Writing assessment has a long history, and one could be forgiven for thinking that we know a lot about what writing is, how it develops, and how to assess it. But as with many issues in language testing, the reality is more complicated. Writing is a constantly moving target for assessment; literacy practices evolve, and writing technologies change. In Chapter 14, Knoch carefully charts the different methods for conceptualising a writing construct and discusses a range of challenges related to scoring performances and interpreting the meaning of results. As with other chapters, technology looms large in new methods of automated scoring and feedback provision and in newer affordances raised by technology, such as the increasing prevalence of collaborative writing on platforms such as Google Docs. One of the most interesting contributions in this chapter, though, is conceptual. Knoch draws together research on written corrective feedback and writing assessment, arguing for greater merging of work in these two areas. As the field increasingly turns its attention to learning-oriented approaches to assessment, we might anticipate that feedback will become a central concern in writing assessment, even for large-scale international exams.

Based firmly in reading theory and the history of psycholinguistic research into reading processes, Brunfaut offers a masterly survey of approaches to the assessment of second language reading over the decades. Covering reading in both a first and a second language, she traces the evolution of reading assessment from the early days to the present. The discussion is illustrated with well-chosen references to some of the most influential reading tests currently in use. The summary of research into reading assessment is comprehensive and impressive and thoroughly supports the description of a range of research methodologies that practitioners can use in both creating and investigating the validity of reading tests. Brunfaut concludes Chapter 15 with directions for future research, which should provide an excellent starting point for anyone wishing to develop a reading assessment project that would add to our understanding of the field.

Section 5 Test design and administration

In Chapter 16 on test specifications (specs), Yan Jin provides a description of the role and purpose of test blueprint documents for both high- and low-stakes tests. The piercing analysis reveals the critical role of specs within the two paradigms: specifying purpose, content, and structure, and creating parallel forms. In the context of low-stakes assessment, their role in defining curriculum content and enhancing local understanding of learning goals is explained with clarity. The role of specs is supported with excellently chosen examples to illustrate their use across contexts. Yan Jin draws upon Fred Davidson's work in the previous edition of this volume as well as that of others but adds to our understanding of the richness of possibility in spec use. Pointing to the future, it is suggested that test specs may also be used to articulate intended consequences. What we like about this chapter is its inclusivity, the mastery over the history of the field, and the foresight offered through such a thorough understanding as well as practical knowledge.

Evidence-centered design (ECD) has increasingly been used in both test design, and the evaluation of test use and retrofit, as the chapter by Yan Jin in this volume attests. In Chapter 17 Yin and Mislevy provide us with a clear definition of ECD and its design components, along with an explanation of how these relate to argument-based approaches and a number of other validity models. It is arguably the flexibility of ECD that has led to its widespread use by test producers in North America, and its modular approach allows reuse of design elements across tests. While ECD can sometimes seem complex, coming to

terms with the design components can provide test designers with a conceptual framework for their activity.

Chapter 18 deals with accommodations to tests and assessments and is an updated version of Abedi's chapter in the first edition of this book. Accommodations are changes or alterations to a test in order to compensate for some disadvantage that may lower test scores as a result of a construct-irrelevant disability. These accommodations may compensate for problems such as a hearing or sight impediment or, in the case of language learners taking content tests (e.g., mathematics or history), provide them with access to the material without language ability impacting the test score. It is arguably the case that this is where validity theory meets practice head on. The reason is that any accommodation should reduce the impact of construct-irrelevant factors (the disability or taking the test in a second language) so that the score reflects ability on the construct; yet, if the accommodation alters the definition of the construct, as would be the case if a text in a reading comprehension test were read aloud to a blind test taker, the score meaning is fundamentally altered. Secondly, if the accommodation would raise the scores of test takers who would not normally be eligible for an accommodation, questions may be raised regarding the "fairness" of the practice. This may be the case with allowing additional time to dyslexic candidates, for example, if the same accommodation had a similar impact on scores of non-dyslexic candidates. Abedi also refers to the concept of universal design, which is becoming increasingly important to create accessible materials and avoid litigation on the grounds of discrimination.

Rater variability is a perennial topic in language testing research. As long as there are human raters, there will be interest in rater cognition, rater behaviour, and rater bias. The results of such studies are often fascinating, but they also lead to one inevitable conclusion: the need for rater training. Variability is seen as a threat to fairness. But how do we train effectively? Is training worth the effort? What does being an expert rater entail? And is variability actually all that bad? Davis provides a comprehensive overview of what we currently know about rater training (as well as interlocutor training, relevant to those tests with examiner-interlocutors), pointing to evidence-based findings and best-practice approaches. One of the most intriguing parts of Chapter 19 is the acknowledgement that strong uniformity is not always desirable. To some extent, training raters to be interchangeable with other raters only paves the way for automated scoring systems, which do uniformity much better and at a far lower cost. At the same time, we would not want to see a return to a system such as that practiced in the original Cambridge Proficiency in English exam, in which – as described by Davis – reputable individuals made judgments of the acceptability of language according to personal taste. Nevertheless, we believe there is always room for a human approach to language assessment, and what is more human than variability? The best kind of training, then, might not be the one that brings all raters into line, but the one which allows raters to form a community: to share their interpretations of rubrics and to increase their understanding of the nature of the construct.

Section 6 Writing items and tasks

There is a very serious scarcity of research and scholarship in our field about one of the most fundamental activities of language testing: item writing. While there are signs that this is changing, particularly as graduate student projects gravitate toward this obvious gap in the literature, there remain some difficult challenges in conducting research on item writing practices: security, ethics, and the general reluctance of item writers to share their "real" experiences on the job for fear of saying the wrong thing. In Chapter 20, Shin addresses this

point but takes the issue much further. Shin effectively opens up a range of possibilities for further research into item writing and item writers, pointing both to the micro-level issues at the heart of the item writing process and also the broader socio-political issues which guide the creation or selection of items for test construction. One of the most vivid passages in this chapter posits that item writers can be both oppressed (through the precarity of their working practices) and oppressors (possessing considerable power to select and write items which are then presented to test takers). In passages such as this, Shin's chapter lifts the study of item writing practices beyond the more mundane questions of test assembly and creates a new focus for language testing research: the item writer as a complex social actor and item writing as an important site of critical inquiry.

For a long time, it has been argued that an item or task should test a particular ability or skill in isolation so that scores are not contaminated. By "contamination," the critics meant (for example) that an ability to read should not cause a score on a speaking or writing test to vary. Recently, however, there has been a renewed interest in integrated items. The rationale is that such items more precisely reflect the kinds of holistic tasks that are undertaken in the "real world." Students do not simply write essays from prompts. They read texts, listen to lectures, discuss ideas, and then come to the writing task. In this revised Chapter 21, Plakans addresses the issues of complex constructs, how they are operationalized in integrated tasks, and the problems surrounding score interpretation. She provides new illustrations of integrated task types and updates us on the important research that has been carried out in recent years. The evolution and increased adoption of these task types will make integrated assessment a prolific area of research for many years to come.

In Chapter 22, Cohen updates his cutting-edge discussion of test-taker strategies. Central to the discussion is how, and to what extent, strategy use is part of the construct being assessed or something that fundamentally undermines our ability to make meaningful inferences about ability from test responses and scores. In the descriptive and historical parts of the chapter, Cohen adds to what we have learned about test-taking strategies, including the *social* element of their nature. He adds extensively to the list of new studies on test-taking strategies, providing an overview of advances in the field. The major new contribution to the debate comes with the new concept *test-deviousness strategies* and the role of these strategies in subverting valid score interpretation. This is a more accurate term than *test-wiseness* for strategies designed to result in higher scores without a corresponding increase in the construct of interest. In debunking much of what has proved less than useful in strategy research and setting us on a firmer path, Cohen has produced a chapter that is essential reading for anyone planning to undertake research in this area.

Section 7 Prototyping and field tests

Language test development has many commonalities with other areas of human endeavour that involve design. There is first an idea or thinking stage, followed by a stage in which those ideas are made more concrete through plans, specifications, and models. Developing a prototype is a fundamental step – often, this is the first point at which an idea is made tangible: an important means of communicating the design to a wider audience. In language testing, prototyping can often be a challenging step in the design process – full of trial and error – but it is also the point at which language test development can be exciting, as a new task comes to fruition, and theoretical potential is realised. In Chapter 23, Nissan and Park provide a clear overview of how these processes work at Educational Testing Service, giving some real examples of prototyping and talking through the various methods involved in

determining how to revise a task and whether to take it through to the next stage. As with other chapters in this collection, Nissan and Park point toward the increasing importance of digital technology in prototyping. This chapter will be of great interest to anyone involved in test design or anyone who wants to understand the role of design in tests.

Chapter 24 by Kremmel, Eberharter, and Holzknecht replaces and updates Chapter 20 in the first edition of the *Handbook* by Kenyon and MacGregor. They have ensured that the processes and practices so successfully described in the first chapter are still present, but they have added to the discussion with their own test development experience in Europe. This chapter, like the one before it, is essential reading for test developers. The simple reason, as the authors explain, is that there is a dearth of literature on pre-operational processes that must be carried out before a test is used to make important decisions. All too often, tests and assessments are written quickly and go straight into operational use without all the quality controls that this chapter so adeptly describes. Together with the contribution of Nissan and Park, we are provided with comprehensive test development guidance.

Chapter 25 – written by John Read – retains the dual focus that it had in the first edition of the *Handbook*. On the one hand, it is about the practice and process of piloting tests and so complements the new chapter by Kremmel *et al*., and the updated chapter by Nissan and Park. However, it does this from the perspective of vocabulary testing. Vocabulary testing continues to evolve as a very distinct discipline within language testing, and the more widespread use of corpora (see Cushing, this volume) has only added to the number of vocabulary studies being conducted. Vocabulary tests remain widely used as placement instruments and for research into how the brain stores lexical items and meaning. It is therefore crucial that these tests are piloted with care, and as John Read observes, the literature really does not tell us how piloting is carried out. With reference to the word associates format, he guides the reader through the process of piloting a vocabulary test and, in the process, provides us with a generic template for piloting other tests as well. The new chapter is updated with current literature, making it an invaluable resource for all vocabulary test researchers.

Section 8 Measurement theory in language testing

Perhaps the oldest psychometric toolkit available to language testing is CTT, or classical test theory. In Chapter 26, J. D. Brown reminds us that these tools are not obsolete. CTT is in widespread and perhaps pervasive use because it provides test developers with a means to carefully monitor the contribution of each test item to the overall test score distribution. The assumptions underlying modern test theory are essentially the same as those in CTT, and it is arguably the case that CTT offers insights that we have lost in some of the statistical determinism that can accompany more recent innovations in statistical analysis. This treatment of the "classical" toolbox is not only highly relevant to current practice; it is also essential for all language testers to gain an understanding of how we use statistical analysis in language testing research and design.

Item response theory (IRT) and many-facet Rasch measurement (MFRM) are now commonplace methods in language testing research. Aspiring testers are encouraged to learn these techniques in order both to be able to use these statistics in their research and development work and to develop a critical statistical literacy for reading and evaluating the research literature of the field. For those with little background in these methods, a daunting question is often "Where do I start?" Ockey's updated Chapter 27 in the second edition of the *Handbook* is a model of clarity. This is a perfect entry point for anyone interested in the

fundamental concepts in IRT and MFRM and provides an excellent discussion of issues for those already familiar with these approaches. Ockey includes coverage of recent literature and provides direction for interested readers to continue exploring the area. We encourage anyone with an interest in IRT and MFRM to dip into Ockey's chapter for an up-to-date overview.

In a similar way to Ockey's chapter, Chapter 28 on reliability by Yan and Fan renders a topic difficult to understand into a model of clarity. Their treatment of essential concepts in all measurement is completely accessible, and so when they go on to look at the five commonly used reliability coefficients, the reader is able to understand the type of measurement error that each one addresses with ease. But the chapter is not just a comprehensible overview of reliability. The authors offer new insights into the theoretical place of reliability within language testing and educational measurement with a nuanced treatment of its role and meaning in measurement paradigms. Far too often, reliability is treated simply as the application of psychometric technique. The authors' understanding of this topic is far too deep for such a fallacy to remain unchallenged. This chapter is set to become a classic in the field and should be read by researchers in educational measurement and behavioural science research more generally.

The Measurement section is rounded off by Chapter 29, in which Galaczi and Lim provide a thoughtful consideration of the very human factors that go into making judgments about performance. Notwithstanding current research into automated scoring (see the chapter by Xi, as well as the discussion here), it remains the case that communication takes place between complex biological beings, and deciding whether that communication is successful for a particular purpose is a task best suited to human inference. The authors set out some of the threats to sound inference and agreement between humans and describe the range of scoring options that have been devised to address concerns. But they argue, correctly, that there is no single solution and that there is no such thing as a scoring mechanism that is universally "good." Nevertheless, the research that has led us to where we are allows Galaczi and Lim to set out suggestions for practice that will be of benefit to all who are tasked with designing a scoring system for a performance test.

Section 9 Technology in language testing

Automated scoring systems are on the rise in language testing, extending from their use in high-stakes testing to more low-stakes classroom applications (e.g., automated writing evaluators [AWEs]). Their utility is obvious: they can bypass the need for human raters, allowing for the processing of large volumes of speaking and writing performances, at great speed and with a very quick turnaround for results. This is not just an economic argument; human raters bring with them biases that can impact scoring, and automated systems can maintain consistency in a way that is difficult for humans. Automated scoring, though, brings unique challenges with respect to validity. As Xi puts it in Chapter 30, "an automated scoring system is not just a case of replacing the human rater; rather, the system interacts with the other assessment components in complex ways." In her updated chapter, Xi outlines the validity issues associated with automated scoring, highlighting key questions that need to be addressed in supporting their use in operational testing. Xi also raises a fascinating discussion point: what sort of AI (artificial intelligence) literacy is required for stakeholders to engage with automated scoring systems, to understand their workings, and to critique their design? Relating to this point is the general issue of transparency. It remains to be seen how producers of automated scoring systems will grapple with balancing transparency and

proprietary knowledge. In the interests of stakeholders, we hope that the balance is toward the former.

Working through the chapters that were revised or newly written for this second edition, there was one that stood out as a common cross-reference: computer-based testing by Sawaki. Digital technology and computer-based testing are major themes across the whole collection. This is not necessarily related to the influence of the pandemic; most first drafts of chapters were produced prior to its onset, demonstrating that technology was already a major concern for language testers, weaving its way into a range of topics. However, the pandemic has increased the relevance of the focus on computer-based testing and amplified many of the issues that were already in train. Against this backdrop, Sawaki was given the unenviable task of encapsulating the current state of play in just one chapter. We think she has done a truly admirable job. Chapter 31 moves through the historical trajectory of computer-based testing before tackling some current issues. There is coverage here of more conventional comparative research (computer-based versus paper-based), as well as the novel and innovative uses of technology through virtual reality and spoken dialog systems. Similar to other authors, Sawaki sees potential for technology to connect assessment with learning more effectively. The chapter ends with a call for more collaboration with others outside the field to best harness the potential of technology. We agree with this sentiment and will look forward to future interdisciplinary work that explores computer-based testing from all angles: educational, technical, social, and ethical.

In deciding on new chapters for the second edition of the *Handbook*, one non-negotiable inclusion was a chapter on corpus linguistics and language testing. In the past 10 to 15 years, corpus approaches have become so commonplace in the practices of language testers that, as Cushing says in this chapter, "it would be almost unthinkable now to develop a large-scale language test without referring to corpus data." There are various ways in which corpora can be used in language testing. In Chapter 32, Cushing explains very clearly the various common methods of corpus linguistics and demonstrates the utility of corpus approaches for language testers. Some of the key applications include describing the domain of language use, determining features of performance across levels, and providing evidence to support extrapolation inferences. Corpora are not without their limitations, though, and language testers need to become not only corpus conversant, but also corpus critical. Cushing further points out that corpus techniques are closely interconnected with automated scoring systems. For that reason, language testers are encouraged to work together with computational linguists to create better corpora to inform scoring models.

Section 10 Ethics, Fairness, and Policy

The updated Chapter 33 by Walters incorporates fairness and ethics, which were separate chapters in the first edition of the *Handbook*. Building on the previous chapter by Davies, Walters incorporates a discussion and definition of the various ethical positions that language testers have adopted. The discussion of fairness then sits within the framework established. The chapter also treats a currently unresolved philosophical question: how is fairness different from validity? To close the chapter, Walters still leaves the reader with several carefully constructed exercises that help them relate the concepts and frameworks to their own assessment contexts.

Standards are endemic in all walks of life. Language education and testing have attempted to emulate what happens in industry. While analogy is a tenuous basis for a rationale, the fact is that creating and promoting language standards has become an industry. In this

masterly review and analysis, Deygers classifies standards into three types: educational performance indicators, proficiency frameworks, and institutionalized language tests. Chapter 34 describes each in turn and presents us with use, research, and critique. This is an area that cuts across policy, theory, research, and practice. So it is not surprising that it is highly controversial. Deygers is a sure guide to potential pitfalls and opportunities.

Language tests are developed across a variety of contexts. Many well-known tests are produced by large commercial or nonprofit organisations, with teams of staff working on assessment development, rating, research, and marketing. Other tests are developed by small teams of teachers with little support from their institutions. Each case presents different challenges for maintaining quality. In the case of smaller teams, it may be a lack of resources (both material and human) with few opportunities for conducting post hoc analysis of results. For larger providers, it is the complexity of the organisation itself: keeping track of who is doing what, ensuring communications are clear, avoiding "silo thinking," and so on. Saville and McElwee provide an updated version of an important chapter focusing on quality management in language testing. Combining the approaches of management science with language testing's typical focus on validity, Chapter 35 provides a useful template for test developers of all sizes to consider their own quality control procedures.

In the final chapter, we provide an epilogue to the volume as a whole. The first edition of the *Handbook* did not contain an epilogue; Fulcher and Davidson were satisfied to let the chapters have the final say. We maintain this stance, and our epilogue is not intended to critique or question the arguments put forth in the collection. Rather, we wanted to use the epilogue to reflect on themes and issues that we noticed emerging across the collection of chapters from our vantage point as editors. Or, to use a musical analogy, we wanted to riff on the multi-layered tracks laid down by our collaborators. From this perspective, we identify four key themes: (1) the increasing role of technology, (2) connecting assessment with learning, (3) grappling with complexity, and (4) theorising the socio-political nature of language testing. And having identified and discussed these themes, we offer some predictions for the direction of travel in the field as a whole. We end the epilogue by reiterating a call made by Fulcher and Davidson (2007) for a shift toward effect-driven testing. Understanding the scope and nature of the field, however, is the necessary precursor for predicting such effects. A volume of this kind tells us where we have been and where we need to go.

References

Aurelius, M. (2013). *Marcus Aurelius: Meditations, Books 1–6*. Oxford: Oxford University Press.

Bachman, L. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville (ed.), *Inference and Generalizability in Applied Linguistics: Multiple Perspectives*. Amsterdam: John Benjamins, 165–207.

Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Mill, J. S. (1859). On Liberty. In J. Gray (ed.), *John Stuart Mill On Liberty and Other Essays*. Oxford: Oxford University Press.

Editorial

Aurelius, M. (2013). Marcus Aurelius: Meditations, Books 1–6. Oxford: Oxford University Press.
Bachman, L. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville (ed.), Inference and Generalizability in Applied Linguistics: Multiple Perspectives. Amsterdam/: John Benjamins, 165–207.
Fulcher, G. (2015). Re-examining Language Testing: A Philosophical and Social Inquiry. London and New York: Routledge.
Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York: Routledge.
Mill, J. S. (1859). On Liberty. In J. Gray (ed.), John Stuart Mill On Liberty and Other Essays. Oxford: Oxford University Press.

Conceptions of validity

Bachman, L. F. and Palmer, A. S. (2010). Language Assessment in Practice. Oxford: Oxford University Press. This book takes readers through the process of building an assessment use argument based on four claims. It is aimed at readers wishing to take validity argumentation into account from the beginning stages of test design without getting into the detail of a more complex validity argument. It emphasizes the use of validity arguments for justifying test design decisions to stakeholders.
Chapelle, C. A. (2021). Argument-based Validation in Testing and Assessment. Thousand Oaks, CA: Sage Publishing. Chapelle defines the concepts in assessment required to develop and use validity arguments by situating them historically and introduces the language required to create and understand complex validity arguments. She illustrates how the basic concepts in assessment are made more precise and contextually relevant by developing them into claims, inferences, warrants, and assumptions in validity arguments.
Chapelle, C. A. , Enright, M. E. and Jamieson, J. (2010). Does an argument-based approach to validity make a difference? Educational Measurement: Issues and Practice 29: 3–13. This paper describes four points of distinction that can be made between the approach to validation described by Messick and the more praxis-oriented argumentation outlined by Kane. It draws upon experience in developing a validity argument for a high-stakes language test that attempted both approaches.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement 50: 1–73. This article supersedes previous presentations of validity argument by Kane. It presents the historical context in which argument-based validity took shape, introduces terms, and presents the concepts of validity argument. This is an important reference for anyone working on a validity argument.
Kunnan, A. J. (2018). Evaluating Language Assessments. New York and London: Routledge. The volume develops the issues entailed in the wider political, economic, social, legal, and ethical contexts in which validity arguments are intended to work. Through the lens of fairness of test use and justice, the book presents the types of qualitative and quantitative research that can be used in validation studies. The book demonstrates the importance of theory and practice in language assessment for the fair use of language tests and achievement of social justice.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Macmillan Publishing Co, 13–103. This is the seminal paper on validity, which consolidates and extends work on theory and practice in validation of educational and psychological assessments while contextualizing validation within philosophy of science. It defines validity in terms of interpretations and uses of tests while including relevance and utility, value implications, and consequences of testing. It contains extensive discussion and examples of validation research in addition to definition and discussion of concepts that remain fundamental conceptions of validity.
AERA, APA and NCME . (2014). Standards for Educational and Psychological Testing, 6th edn. Washington, DC: AERA.
Alderson, J. C. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge, UK: Cambridge University Press.
Anastasi, A. (1954). Psychological Testing. New York, NY: Palgrave Macmillan.
Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford, UK: Oxford University Press.
Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quarterly 2 1–34. https://doi.org/10.1207/s15434311laq0201_1.
Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox , M. Wesche and D. Bayliss (eds.), What Are We Measuring? Language Testing Reconsidered. Ottawa, Canada/: University of Ottawa Press, 41–71.

Bachman, L. F. and Dambock, B. (2017). Language Assessment for Classroom Teachers. Oxford, UK: Oxford University Press.
Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice. Oxford, UK: Oxford University Press.
Bachman, L. F. and Palmer, A. S. (2010). Language Assessment in Practice. Oxford, UK: Oxford University Press.
Brennan, R. L. (ed.), (2006). Perspectives on the evolution and future of educational measurement. In R. Brennen (ed.), Educational Measurement, 4th edn. Westport, CT: Greenwood Publishing, 1–16.
Brooks, L. and Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBTTM and real-life academic speaking activities. Language Assessment Quarterly 11 353–373. https://doi.org/10.1080/15434303.2014.947532.
Byrnes, H. , Maxim, H. H. and Norris, M. J. (2010). Realizing advanced foreign language writing: Development in collegiate education: Curricular design, pedagogy, assessment. Modern Language Journal 94: 1–8.
Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56 81–105. https://doi.org/10.1037/h0046016.
Chalhoub-Deville, M. , Chapelle, C. A. and Duff, P. A. (2006). Inference and Generalizability in Applied Linguistics: Multiple Perspectives. Amsterdam, Netherlands: John Benjamins Publishing Company.
Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), Interfaces between Second Language Acquisition and Language Testing Research. Cambridge, UK/: Cambridge University Press, 32–70.
Chapelle, C. A. (2021). Argument-based Validation in Testing and Assessment. Thousand Oaks, CA: Sage Publishing.
Chapelle, C. A. and Douglas, D. (1993). Foundations and directions for a new decade of language testing. In D. Douglas and C. Chapelle (eds.), A New Decade of Language Testing Research. Arlington, VA/: TESOL Publications, 1–22.
Chapelle, C. A. , Enright, M. E. and Jamieson, J. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language. London, UK: Routledge.
Chapelle, C. A. and Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (ed.), The Companion to Language Assessment. Chichester, UK/: Wiley, 1079–1097.
Cheng, L. and Sun, Y. (2015). Interpreting the impact of the Ontario secondary school literacy test on second language students within an argument-based validation framework. Language Assessment Quarterly 12 50–66. https://doi.org/10.1080/15434303.2014.981334.
Cherryholmes, C. (1988). Power and Criticism: Poststructural Investigations in Education. New York, NY: Teachers College Press.
Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun (eds.), Test Validity. Hillsdale, NJ/: Lawrence Erlbaum, 3–17.
Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin 52 281–302. https://doi.org/10.1037/h0040957.
Cumming, A. (1996). Introduction: the concept of validation in language testing. In A. Cumming and R. Berwick (eds.), Validation in Language Testing. Clevedon, UK/: Multilingual Matters, 1–14.
Davies, A. and Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (ed.), Handbook of Research in Second Language Teaching and Learning. Mahwah, NJ/: Lawrence Erlbaum Associates, 795–813.
Douglas, D. (2000). Assessing Language for Specific Purposes. Cambridge, UK: Cambridge University Press.
Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. Language Testing 27 317–334. https://doi.org/10.1177/0265532210363144.
Ercikan, K. W. and Pellegrino, J. W. (eds.). (2017). Validation of Score Meaning in the Next Generation of Assessments. The Use of Response Processes. New York: Routledge.
Finocchiaro, M. and Sako, S. (1983). Foreign Language Testing: A Practical Approach. New York, NY: Regents Publishing Company.
Fulcher, G. (2015). Re-examining Language Testing: A Philosophical and Social Inquiry. New York, NY: Routledge.
Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London, UK: Routledge.
Halliday, M. A. K. and Hasan, R. (1989). Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective. Oxford, UK: Oxford University Press.
Harris, D. P. (1969). Testing English as a Second Language. New York, NY: McGraw Hill.

Heaton, J. B. (1975). Writing Language Tests. London, UK: Longman.
Hughes, A. (1989). Testing for Language Teachers. Cambridge, UK: Cambridge University Press.
Johnson, R. C. (2011). Assessing the Assessments: Using an Argument-Based Validity Framework to Assess the Validity and Use of an English Placement System in a Foreign Language Context. Unpublished doctoral dissertation, Macquarie University, Sydney, Australia.
Jones, N. and Saville, N. (2016). Learning Oriented Assessment: A Systemic Approach. Studies in Language Testing, vol. 45. Cambridge: UCLES/Cambridge University Press.
Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin 112 527–535. https://doi.org/10.1037/0033-2909.112.3.527.
Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement 38 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x.
Kane, M. T. (2006). Validation. In R. Brennen (ed.), Educational Measurement, 4th edn. Westport, CT: Greenwood Publishing, 17–64.
Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement 50 1–73. https://doi.org/10.1111/jedm.12000.
Kane, M. T. , Crooks, T. and Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice 18 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.
Klebanov, B. B. , Ramineni, C. , Kaufer, D. , Yeoh, P. and Ishizaki, S. (2019). Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. Language Testing 36 125–144. https://doi.org/10.1177/0265532217740752.
Kunnan, A. J. (1998). Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach. Mahwah, NJ: Lawrence Erlbaum Associates.
Kunnan, A. J. (2018). Evaluating Language Assessments. New York and London: Routledge.
Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. New York, NY: McGraw-Hill.
Lissitz, R. W. (ed.). (2009). The Concept of Validity: Revisions New Directions and Applications. Charlotte, NC: Information Age Publishing, Inc.
Llosa, L. (2005). Building and Supporting a Validity Argument for a Standards-Based Classroom Assessment of English Proficiency. Unpublished doctoral dissertation, University of California, Los Angeles.
Llosa, L. and Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. Language Testing 36 235–263. https://doi.org/10.1177/0265532218763456.
Markus, K. A. and Borsboom, D. (2013). Frontiers of Test Validity Theory: Measurement, Causation, and Meaning. New York, NY: Routledge.
McNamara, T. (1996). Measuring Second Language Performance. London, UK: Longman.
McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. Language Assessment Quarterly 3 31–51. https://doi.org/10.1207/s15434311laq0301_3.
McNamara, T. and Roever, C. (2006). Language Testing: The Social Dimension. Malden, MA: Blackwell Publishing.
Messick, S. A. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin 89 575–588. https://doi.org/10.1037/0033-2909.89.3.575.
Messick, S. A. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Macmillan Publishing Co, 13–103.
Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (ed.), The Concept of Validity: Revisions New Directions and Applications. Charlotte, NC/: Information Age Publishing, Inc, 83–108.
Newton, P. E. and Shaw, S. D. (2014). Validity in Educational & Psychological Assessment. London, UK: Sage Publications.
Norris, J. (2008). Validity Evaluation in Foreign Language Assessment. New York, NY: Peter Lang.
Oller, J. (1979). Language Tests at School. London, UK: Longman.
Purser, E. , Dreyfus, S. and Jones, P. (2020). Big ideas & sharp focus: Researching and developing students' academic writing across the disciplines. Journal of English for Academic Purposes 43: 100807. https://doi.org/10.1016/j.jeap.2019.100807.
Sawaki, Y. and Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? Language Testing 35 529–556. https://doi.org/10.1177/0265532217716731.
Shohamy, E. (2001). The Power of Tests: A Critical Perspective on the Uses of Language if Language Tests. Harlow, UK: Pearson Education.
Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (ed.), The Concept of Validity: Revisions New Directions and Applications. Charlotte, NC/: Information Age Publishing, Inc, 19–38.

Sireci, S. and Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema 26 100–107. https://doi.org/10.7334/psicothema2013.256.
Stoynoff, S. and Chapelle, C. A. (2005). ESOL Tests and Testing: A Resource for Teachers and Program Administrators. Alexandria, VA: TESOL Publications.
Valette, R. A. (1967). Modern Language Testing. New York, NY: Harcourt, Brace and World.
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. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa, USA.
Wang, H. , Choi, I. , Schmidgall, J. and Bachman, L. F. (2012). Review of Pearson Test of English Academic. Language Testing 29 603–619. https://doi.org/10.1177/0265532212448619.
Weir, C. (2005). Language Testing and Validation: An Evidence-Based Approach. Hampshire, UK: Palgrave Macmillan.
Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. Language Testing 32 199–225. https://doi.org/10.1177/0265532214557113.
Zwick, R. (2006). Higher education admissions testing. In Educational Measurement, 4th edn. Westport, CT: Greenwood Publishing, 221–256.

Articulating a validity argument

Bachman, L. (2005). Building and supporting a case for test use. Language Assessment Quarterly 2: 1–34. Bachman (2005) provides a very clear and concise introduction to an argument-based framework for validation, with an emphasis on the evaluation of test uses in terms of the consequences of such use. He examines the kinds of warrants and evidence needed for a utilization argument over and above those typically needed for the validations of a score interpretation per se.
Chapelle, C. A. , Enright, M. and Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright and J. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language. New York, NY: Routledge, 1–25. Most of the examples of IUAs and validity arguments in the literature are quite simple and quite general and have been developed for illustrative purposes. Chapelle, Enright, and Jamieson (2004) developed and documented an IUA and a validity argument for a large and complex testing program and, in doing so, have pushed the argument-based methodology forward in a dramatic way by applying the approach to an existing assessment program.
Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin 52: 281–302. It took a while for the sophisticated framework introduced by Cronbach and Meehl to gain traction, but it became the base for most work on the theory of validity from the 1980s to the present. It introduced a conceptualization of score interpretation and validation that was much broader and more sophisticated than those that had been in use up to that point. This article is over 50 years old and does not provide specific guidance for validation, but it is essential reading for anyone who wants to follow current debates on validity and validation.
Kane, M. T. (2006). Validation. In R. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: American Council on Education and Praeger, 17–64. Kane provides an overview of an argument-based approach to validation and indicates how some common interpretations and uses of test scores can be validated.
Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement 50: 1–73. Kane (2013) provides an updated version of the argument-based validation described in Kane (2006), with greater emphasis on validating score uses.
American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, and NCME) . (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.
Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quarterly 2 1–34. https://doi.org/10.1207/s15434311laq0201_1.
Bachman, L. F. and Palmer, A. (2010). Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World. Oxford, UK: Oxford University Press.
Brennan, R. (2001). Generalizability Theory. New York, NY: Springer-Verlag.
Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. Lissitz (ed.), The Concept of Validity. Charlotte NC/: Information Age Publishing, 241–263.
Chalhoub-Deville, M. and Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: Praeger Publishers, 517–530.
Chapelle, C. A. (1999). Validity in language assessment. Annual Review of Applied Linguistics 19 254–272. https://doi.org/10.1017/S0267190599190135.

Chapelle, C. A. , Enright, M. and Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle , M. Enright and J. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language. New York, NY/: Routledge, 1–25.
Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.), Educational Measurement, 2nd edn. Washington, DC: American Council on Education, 443–507.
Cronbach, L. J. (1980). Validity on parole: How can we go straight? In New Directions for Testing and Measurement: Measuring Achievement Over a Decade, 5. San Francisco, CA/: Jossey-Bass, 99–108.
Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun (eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum, 3–17.
Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin 52 281–302. https://doi.org/10.1037/h0040957.
Frisbie, D. A. (1982). Methods of evaluating course placement systems. Educational Evaluation and Policy Analyses 4 133–140. https://doi.org/10.3102/01623737004002133.
Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London, UK: Routledge.
Kane, M. T. (2006). Validation. In R. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: American Council on Education and Praeger, 17–64.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement 50 1–73. https://doi.org/10.1111/jedm.12000.
Kane, M. T. , Crooks, T. J. and Cohen, A. S. (1999). Validating measures of performance. Educational Measurement: Issues and Practice 18 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: American Council on Education and Macmillan, 13–103.
Toulmin, S. (1958). The Uses of Argument. Cambridge, UK: Cambridge University Press.
Widdowson, H. G. (2001). Communicative language testing: The art of the possible. In C. Elder , A. Brown , E. Grove et al. (eds.), Experimenting with Uncertainty, Essays in Honor of Alan Davies. Cambridge, UK/: Cambridge University Press, 12–21.
Xi, X. (2008). Methods of test validation. In E. Shohamy and N. Hornberger (eds.), Encyclopedia of Language and Education: Language Testing and Assessment, 2nd edn, Vol. 7. New York, NY: Springer, 177–196.

Inference and prediction in language testing

Lipton, P. (2004). Inference to the Best Explanation, 2nd edn. London: Routledge. Peter Lipton summarizes the history of causal inference from a philosophy of science perspective. Lipton also provides an overview of how hypotheses evolve and how evidence is considered in light of plausible alternatives using observational, experimental, and Bayesian methods of interpretation.
Mislevy, R. J. (2017). Sociocognitive Foundations of Educational Measurement. New York: Routledge. In this comprehensive text, Robert J. Mislevy lays out an approach to integrating measurement theory, learner experiences, and contextual factors into the interpretation of measurement outcomes in an evidence-centered design framework.
Morgan, S. and Winship, C. (2007). Counterfactuals and Causal Inference: Methods and Principles for Social Research. New York: Cambridge University Press. Morgan and Winship outline the quantitative methodology needed to untangle the complexities of direct and indirect causal arguments, especially in non-experimental and ex-post-facto designs.
Singer, J. and Willett, J. (2003). Applied Longitudinal Data Analysis. New York: Oxford University Press. Singer and Willett provide an accessible and comprehensive overview of quantitative methods used in longitudinal data analysis, including multi-level modeling, latent growth curve models, and event history analysis.
Asher, H. B. (1983). Causal Modeling. Sage University Paper series on Quantitative Methods in the Social Sciences. Beverly Hills and London: Sage Press.
Bachman, L. (2005). Building and supporting a case for test use. Language Assessment Quarterly 2 1–34. https://doi.org/10.1207/s15434311laq0201_1.
Baum, S. and Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. Applied Psycholinguistics 35 857–894. https://doi.org/10.1017/S0142716414000174.
Bernstein, J. , van Moere, A. and Cheng, J. (2010). Validating automated speaking tests. Language Testing 27 355–377. https://doi.org/10.1177/0265532210364404.
Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. Journal of the American Statistical Association 67 364–366. https://doi.org/10.1080/01621459.1972.10482387.

Bridgeman, B. , Cho, Y. and Depietro, S. (2016). Predicting grades from English language assessment: The importance of peeling the onion. Language Testing 33 307–318. https://doi.org/10.1177/0265532215583066.
Carroll, J. and Sapon, S. (1959). Modern Language Aptitude Test. New York: American Psychological Corporation.
Chapelle, C. , Enright, M. and Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle , M. Enright and J. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language. New York: Routledge.
Chapelle, C. (2020). Argument-based Validation in Testing and Assessment. London: Sage.
Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. Language Testing 33 453–472. https://doi.org/10.1177/0265532215593312.
Cho, Y. and Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. Language Testing 29 421–442. https://doi.org/10.1177/0265532211430368.
Davies, J. A. (1985). The Logic of Causal Order. Sage University Paper series on Quantitative Methods in the Social Sciences. Beverly Hills and London: Sage Press.
DeKeyser, R. and Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkle (ed.) Handbook of Research in Second Language Teaching and Learning. New York: Routledge.
Doughty, C. J. (2018). Cognitive language aptitude. Language Learning 69 101–126. https://doi.org/10.1111/lang.12322.
Friedman, N. , Miyake, A. , Young, S. DeFries, J. , Corley, R. and Hewitt, J. (2008). Individual differences in executive functions are almost entirely genetic in origin. Journal of Experimental Psychology: General 137 201–225. https://doi.org/10.1037/0096-3445.137.2.201.
Gardner, R. C. (2000). Correlation, causation, motivation, and second language acquisition. Canadian Psychology 41 1–23. https://doi.org/10.1037/h0086854.
Ginther, A. and Yan, X. (2017). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. Language Testing 35 271–295. https://doi.org/10.1177/0265532217704010.
Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In M. Long and G. Granena (eds.), Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment. Amsterdam/: John Benjamins, 105–130.
Hancock, G. R. and Mueller, R. D. (eds.). (2006). Structural Equation Modeling: A Second Course. Greenwich, CT: Information Age Publishing.
Hulstijn, J. H. and Schoonen, R. (2011). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common European framework of reference (CEFR). Language Testing 29 203–221. https://doi.org/10.1177/0265532211419826.
Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: American Council of Education and Praeger Series on Higher Education, 17–64.
Kane, M. , Crookes, T. and Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice 18 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.
Kieffer, M. J. and Lasaux, N. K. (2012). Development of morphological awareness and vocabulary knowledge in Spanish-speaking language minority learners: A parallel process latent growth curve model. Applied Psycholinguistics 33 23–54. https://doi.org/10.1017/S0142716411000099.
Klein, R. (2013). Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, 2nd edn. Washington, DC: American Psychological Association.
Linck, J. A. , Hughes, M. M. , Campbell, S. G. , Silbert, N. H. , Tare, M. , Jackson, S. R. , Smith, B. K. , Bunting, M. F. and Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high level language proficiency. Language Learning 63 530–566. https://doi.org/10.1111/lang.12011.
Lipton, P. (2004). Inference to the Best Explanation, 2nd edn. London: Routledge.
Lissitz, R. and Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher 36 437–448. https://doi.org/10.3102/0013189X07311286.
Llosa, L. (2007). Validating a standards-based classroom-based assessment of English proficiency: A multitrait-multimethod approach. Language Testing 24 489–515. https://doi.org/10.1177/0265532207080770.
Long, M. , Gor, K. and Jackson, S. (2012). Linguistic correlates of second language proficiency: Proof of concept with ILR 2–3 in Russian. Studies in Second Language Acquisition 34 99–126. https://doi.org/10.1017/S0272263111000519.
Luk, G. , Bialystok, E. , Craik, F. I. M. and Grady, C. L. (2010). Lifelong bilingualism maintains white matter integrity in older adults. Journal of Neuroscience 31(46) 16808–16813. https://doi.org/10.1523/JNEUROSCI.4563-11.2011.
Mislevy, R. , Steinberg, L. S. and Almond, R. G. (2002). Design and analysis in task-based language assessment. Language Testing 19 447–496. https://doi.org/10.1191/0265532202lt241oa.

Mislevy, R. (2017). Sociocognitive Foundations of Educational Measurement. New York: Routledge.
Morgan, S. and Winship, C. (2007). Counterfactuals and Causal Inference: Methods and Principles for Social Research. New York: Cambridge University Press.
Morton, J. (2014). Sunny review casts a foreboding shadow over status quo bilingual advantage research. Applied Psycholinguistics 35 929–931. https://10.1017/S0142716414000277.
Noels, K. A. (2001). Learning Spanish as a second language: Learners' orientations and perceptions of their teachers' communication style. Language Learning 51 107–144. https://doi.org/10.1111/0023-8333.00149.
Owen, A. , Hampshire, A. , Grahn, J. , Stenton, R. , Dajani, S. and Burns, A. (2010). Putting brain training to the test. Nature 465 775–776. https://doi.org/10.1038/nature09042.
Petersen, C. and Al-Haik, A. (1976). The development of the defense language aptitude battery (DLAB). Educational and Psychological Measurement 36 369–380. https://doi.org/10.1177/001316447603600216.
Purpura, J. E. , Brown, J. D. and Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research1. Language Learning 65 37–75. https://doi.org/10.1111/lang.12112.
Redick, T. S. , Shipstead, Z. , Harrision, T. L. , Hicks, K. L. , Fried, D. E. , Hambrick, D. Z. , Kane, M. J. and Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. Journal of Experimental Psychology, General 142 359–379. https://doi.org/10.1037/a0029082.
Ross, S. J. (2005). The impact of assessment method on foreign language proficiency growth. Applied Linguistics 26 317–342. https://doi.org/10.1093/applin/ami011.
Sang, F. , Schmidt, B. , Vollmer, J. , Baumert, J. and Roeder, P. (1986). Models of second language competence: A structural equation approach. Language Testing 3 54–79. https://doi.org/10.1177/026553228600300103.
Singer, J. and Willett, J. (2003). Applied Longitudinal Data Analysis. New York: Oxford University Press.
Skehan, P. (2012). Language aptitude. In S. Gass and A. Mackey (eds.), The Routledge Handbook of Second Language Acquisition. London/: Routledge, 381–395.
Tabachnick, B. and Fidell, L. (2007). Using Multivatriate Statistics, 5th edn. Boston: Allyn and Bacon.
Toulmin, S. E. (2003). The Uses of Argument. Cambridge: Cambridge University Press.

Social dimensions of language testing

Davies, A. (ed.). (1997). Ethics in language testing. Special Issue of Language Testing 14: 3. The articles in this special issue were presented in a symposium on the ethics of language testing held at the World Congress of Applied Linguistics in 1996. In ten articles, well-known scholars of language testing address the role of ethics (and the limits of that role) in professional activities such as language testing. The authors discuss language testing as a means of political control, the definition of the test construct, the effects of language tests on the various stakeholders who are involved, and criteria for promoting ethicality in language testing.
McNamara, T. F. and Roever, C. (2006). Language Testing: The Social Dimension. Malden, MA: Blackwell. This book focuses on the social aspects of language testing, including assessment of socially situated language use and societal consequences of language tests. The authors argue that traditional approaches to ensuring fairness in tests go some way to addressing social concerns, but a broader perspective is necessary to understand the functions of tests on a societal scale. They consider these issues in relation to language assessment in oral proficiency interviews and to the assessment of second language pragmatics. They argue that traditional approaches to ensuring social fairness in tests go some way to addressing social concerns, but a broader perspective is necessary to fully understand the social dimensions of language testing.
Shohamy, E. (2006). Language Policy: Hidden Agendas and New Approaches. London, UK: Routledge. Shohamy illuminates the decisions surrounding language policy and tests and emphasizes the effects of these decisions on different groups within society. Drawing on examples from the United States, Israel, and the UK, Shohamy demonstrates different categories of language policy, from explicit use by government bodies and the media to implicit use where no active decisions are made. She also reveals and examines the mechanisms used to introduce language policy, such as propaganda and even educational material. Her critical exploration of language policy concludes with arguments for a more democratic and open approach to language policy and testing, suggesting strategies for resistance and ways to protect the linguistic rights of individuals and groups.
Young, R. F. (2009). Discursive Practice in Language Learning and Teaching. Malden, MA: Wiley- Blackwell. Young sets out to explain practice theory and its implications for language learning, teaching, and testing. He examines the consequences of considering language in interaction as discursive practice and of discourse as social action. Discursive practice is the construction and reflection of social realities through language and actions that invoke identity, ideology, belief, and power. The ultimate aim of practice theory is to explain the

ways in which the global context affects the local employment of communicative resources and vice versa. In Chapters 5 and 6, Young uses practice theory to take a new look at how the employment of communicative resources in a specific discursive practice may be learned, taught, and assessed.

Agar, M. (1994). Language Shock: Understanding the Culture of Conversation. New York, NY: Morrow.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Beyer, C. (2007). Edmund Husserl. In Stanford Encyclopedia of Philosophy. Retrieved September 18, 2010, from plato.stanford.edu/entries/husserl.

Bourdieu, P. (1977). Outline of a Theory of Practice (R. Nice , trans.). Cambridge, UK: Cambridge University Press.

Bourdieu, P. (1990). The Logic of Practice (R. Nice , trans.). Stanford, CA: Stanford University Press.

Bucholtz, M. and Hall, K. (2004). Language and identity. In A. Duranti (ed.), A Companion to Linguistic Anthropology. Malden, MA: Blackwell.

Canagarajah, S. (2009). The plurilingual tradition and the English language in South Asia. AILA Review 22 5–22. https://doi.org/10.1075/aila.22.02can.

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

Chalhoub-Deville, M. and Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (ed.), Handbook of Research in Second Language Teaching and Learning. Mahwah, NJ: Erlbaum.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), Interfaces Between Second Language Acquisition and Language Testing Research. Cambridge, UK: Cambridge University Press.

Chomsky, N. (1986). Knowledge of Language: Its Nature, Origin, and Use. New York, NY: Praeger.

Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge, UK: Cambridge University Press.

de Certeau, M. (1984). The Practice of Everyday Life (S. Rendall , trans.). Berkeley, CA: University of California Press.

Erickson, F. (2004). Talk and Social Theory: Ecologies of Speaking and Listening in Everyday Life. Cambridge, UK: Polity.

Evans, B. A. and Hornberger, N. H. (2005). No child left behind: Repealing and unpeeling federal language education policy in the United States. Language Policy 4 87–106. https://doi.org/10.1007/s10993-004-6566-2.

Ford, C. E. and Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs , E. A. Schegloff and S. A. Thompson (eds.), Interaction and Grammar. Cambridge, UK Cambridge University Press.

Foucault, M. (1978). The History of Sexuality, vol. 1 (R. Hurley , trans.). New York, NY: Pantheon.

Foucault, M. (1995). Discipline and Punish: The Birth of the Prison, 2nd edn. (A. Sheridan , trans.). New York, NY: Vintage.

Foucault, M. and Gordon, C. (1980). Power/Knowledge: Selected Interviews and Other Writings, 1972–1977 (C. Gordon et al., trans.). New York, NY: Pantheon.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. Language Assessment Quarterly 1 253–266. https://doi.org/10.1207/s15434311laq0104_4.

Foucault, M. (2009). Test use and political philosophy. Annual Review of Applied Linguistics 29 3–20. https://doi.org/10.1017/S0267190509090023.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment. London, UK: Routledge.

Gerassimenko, O. , Hennoste, T. , Koit, M. and Raabis, A. (2004). Other-Initiated Self-Repairs in Estonian Information Dialogues: Solving Communication Problems in Cooperation. Paper presented at the Association for Computational Linguistics Special Interest Group Workshop on Discourse and Dialogue, Boston, MA, April 30–May 1.

Giddens, A. (1984). The Constitution of Society: Outline of the Theory of Structuration. Berkeley, CA: University of California Press.

Goffman, E. (1979). Footing. Semiotica 25 1–29. https://doi.org/10.1515/semi.1979.25.1-2.1.

Goffman, E. (1981). Forms of Talk. Philadelphia, PA: University of Pennsylvania Press.

Gumperz, J. J. (1982). Discourse Strategies. Cambridge, UK: Cambridge University Press.

Gumperz, J. J. (1992). Contextualization and understanding. In A. Duranti and C. Goodwin (eds.), Rethinking Context: Language as an In1teractive Phenomenon. Cambridge, UK: Cambridge University Press.

Gumperz, J. J. (1995). Mutual inferencing in conversation. In I. Markova , C. Graumann and K. Foppa (eds.), Mutualities in Dialogue. Cambridge, UK: Cambridge University Press.

Halliday, M. A. K. (1994). Systemic theory. In R. E. Asher and J. M. Y. Simpson (eds.), The Encyclopedia of Language and Linguistics, vol. 8. Oxford, UK: Pergamon.

Kramsch, C. (1986). From language proficiency to interactional competence. The Modern Language Journal 70 366–372. https://doi.org/10.2307/326815.
Lantolf, J. P. (2006). Re(de)fining language proficiency in light of the concept of languagculture. In H. Byrnes (ed.), Advanced Language Learning: The contribution of Halliday and Vygotsky. London, UK: Continuum.
Levinson, S. C. (1992). Activity types and language. In P. Drew and J. Heritage (eds.), Talk at Work: Interaction in Institutional Settings. Cambridge, UK: Cambridge University Press.
Levy, C. J. (2010, June 8). Estonia raises its pencils to help erase Russian. New York Times, A6.
Locke, J. (1690). An Essay Concerning the True Original, Extent, and End of Civil Government. Retrieved September 18, 2010, from http://jim.com/2ndtreat.htm.
Lowenberg, P. H. (1993). Issues of validity in tests of English as a world language: Whose standards? World Englishes 12 95–106. https://doi.org/10.1111/j.1467-971X.1993.tb00011.x.
McNamara, T. F. (1997). Interaction in second language performance assessment: Whose performance? Applied Linguistics 18 446–466. https://doi.org/10.1093/applin/18.4.446.
McNamara, T. F. and Roever, C. (2006). Language Testing: The Social Dimension. Malden, MA: Blackwell.
Mehan, H. (1982). The structure of classroom events and their consequences for student performance. In P. Gilmore and A. A. Glatthorn (eds.), Children in and Out of School: Ethnography and Education. Washington, DC: Center for Applied Linguistics.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: American Council on Education and Macmillan.
Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (ed.), Technical Issues in Large-Scale Performance Assessment. Washington, DC: US Department of Education, Office of Educational Research and Improvement.
Miyazaki, I. (1976). China's Examination Hell: The Civil Service Examinations of Imperial China (C. Schirokauer , trans.). New York, NY: Weatherhill.
Norris, J. M. (2008). Validity Evaluation in Language Assessment. New York, NY: Peter Lang.
O'Brien O'Keeffe, K. (1990). Visible Song: Transitional Literacy in Old English Verse. Cambridge, UK: Cambridge University Press.
Rosenbusch, M. H. (2005). The no child left behind act and teaching and learning languages in U.S. schools. The Modern Language Journal 89 250–261. Retrieved from www.jstor.org/stable/3588685.
Röver, C. (2005). Testing ESL Pragmatics: Development and Validation of a Web-Based Assessment Battery. Frankfurt, Germany: Peter Lang.
Sacks, H. , Schegloff, E. A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language 50 696–735. https://doi.org/10.1016/B978-0-12-623550-0.50008-2.
Sahlins, M. D. (1981). Historical Metaphors and Mythical Realities: Structure in the Early History of the Sandwich Islands Kingdom. Ann Arbor, MI: University of Michigan Press.
Sahlins, M. D. (1985). Islands of History. Chicago, IL: University of Chicago Press.
Saussure, F. de (1983[1916]). Course in General Linguistics (R. Harris , trans.). London, UK: Duckworth.
Schegloff, E. A. and Sacks, H. (1973). Opening up closings. Semiotica 8 289–327. https://doi.org/10.1515/semi.1973.8.4.289.
Schleppegrell, M. J. (2004). The Language of Schooling: A Functional Linguistics Perspective. Mahwah, NJ: Erlbaum.
Seidlhofer, B. , Breiteneder, A. and Pitzl, M.-L. (2006). English as a lingua franca in Europe: Challenges for applied linguists. Annual Review of Applied Linguistics 26 3–34. https://doi.org/10.1017/S026719050600002X.
Shohamy, E. (2001). Democratic assessment as an alternative. Language Testing 18 373–391. https://doi.org/10.1177/026553220101800404.
Shohamy, E. (2004). Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton and K. Toohey (eds.), Critical Pedagogies and Language Learning. Cambridge, UK: Cambridge University Press.
Shohamy, E. (2006). Language Policy: Hidden Agendas and New Approaches. London, UK: Routledge.
Stimpson, G. (1946). A Book About a Thousand Things. New York, NY: Harper.
Trevarthen, C. (1977). Descriptive analyses of infant communicative behaviour. In H. R. Schaffer (ed.), Studies in Mother-Infant Interaction: Proceedings of the Loch Lomond Symposium, Ross Priory, University of Strathclyde, September 1975. London, UK: Academic Press.
Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (ed.), Before Speech. Cambridge, UK: Cambridge University Press.
US Congress . (2002, January 8). No Child Left Behind Act of 2001. Public Law 107–10. United States Congress.
Young, R. F. (2008). Language and Interaction: An Advanced Resource Book. London, UK: Routledge.
Young, R. F. (2009). Discursive Practice in Language Learning and Teaching. Malden, MA: Wiley-Blackwell.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (ed.), Handbook of Research in Second Language Teaching and Learning, vol. 2. London, UK: Routledge.
Young, R. F. and He, A. W. (eds.). (1998). Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency. Amsterdam, The Netherlands: Benjamins.

Designing language tests for specific purposes

Bowles, H. (2012). Analyzing languages for specific purposes discourse. The Modern Language Journal 96: 43–58. This article provides an overview of current methods of discourse-based research in LSP.
Douglas, D. (2000). Assessing Languages for Specific Purposes. Cambridge, UK: Cambridge University Press. This book provides a comprehensive discussion of the issues related to LSP assessment. The author evaluates high-stakes specific-purpose tests by investigating target language use domains and combines insights from applied linguistics with issues related to relevant social/psychological constructs. Specific-purpose tests of listening and speaking, such as the Occupational English Test, the Japanese Language Test for Tour Guides, IELTS, and the Professional Test in English Language for Air Traffic Controllers (PELA) are discussed.
ICAO www.icao.int/icao/en/jr/2004. This link provides online access to English, French, and Spanish versions of the ICAO Journal, an official ICAO publication that includes articles by aviation professionals on issues of current interest. The 2004 Volume 59 Number 1 issue provides background on the ICAO proficiency requirements. The 2008 Volume 63 Number 1 issue gives an update on policy related to member nations' efforts to meet targeted levels.
Knoch, U. and Macqueen, S. (2020). Assessing English for Professional Purposes. Oxon, UK: Routledge. This recent book provides an overview of language assessment for professional purposes (LAPP), touching on healthcare, aviation, and legal contexts, among others.
McNamara, T. and Roever, C. (2006). Language Testing: The Social Dimension. Malden, MA: Wiley-Blackwell. This volume provides a broad-ranging discussion of the need to integrate socio-political considerations into approaches to language testing, with specific reference to test validity, design, and use. Tests of oral proficiency and pragmatics and those used in educational contexts are discussed in detail.
O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. The Modern Language Journal 96: 71–88. This article, part of a focus issue on LSP, gives a thorough history of LSP testing, highlighting the benefits of communication between scholars assessing a variety of languages.
www.aero-lingo.com. This website contains extensive information on the language of air traffic control. It includes a list of 50 airplane crashes connected to language difficulties and related articles. In addition, it provides information about air traffic control phraseology and a useful link to ICAO Doc 9835 (*Manual on the Implementation of ICAO Language Proficiency Requirements*), as well as links to audio files of speech samples of pilots and controllers who have been rated at ICAO Levels 3, 4, and 5.
Alderson, J. C. (2010). A survey of aviation English tests. Language Testing 27 51–72. https://doi.org/10.1177/0265532209347196.
Alderson, J. C. (2011). The politics of aviation English testing. Language Assessment Quarterly 8 386–403. https://doi.org/10.1080/15434303.2011.622017.
Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1: 1–47.
Corpus of Contemporary American English (COCA) . (n.d.). Retrieved from www.americancorpus.org.
Douglas, D. (2000). Assessing Languages for Specific Purposes. Cambridge, UK: Cambridge University Press.
Elder, C. and McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. Language Testing 33 153–174. https://doi.org/10.1177/0265532215607398.
Estival, D. and Molesworth, B. (2009). A study of EL2 pilots radio communication in the general aviation environment. Australian Review of Applied Linguistics 32: 24.1–24.16. https://doi.org/10.2104/aral0924.
Goodwin, M. (1996). Informings and announcements in their environment: Prosody within a multiactivity work setting. In E. Cooper-Kuhlen and M. Selting (eds.), Prosody in Conversation: Interactional Studies. Cambridge, UK/: Cambridge University Press, 436–461.
Hinrich, S. W. (2008). The Use of Questions in International Pilot and Air Traffic Controller Communication. Unpublished doctoral dissertation, Oklahoma State University, Stillwater, OK, USA.
Huhta, A. (2009). An analysis of the quality of English testing for aviation purposes in Finland. Australian Review of Applied Linguistics 32: 26.1–26.14. https://doi.org/10.2104/aral0926.
Hymes, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmes (eds.), Sociolinguistics. Harmondsworth, UK/: Penguin, 269–293.

International Civil Aviation Organization (ICAO) . (n.d.). Strategic Objectives of ICAO. Retrieved September 6, 2009, from www.icao.int/icao/en/strategic_objectives.htm.
International Civil Aviation Organization (ICAO) . (2001). Air Traffic Management Document 4444. ATM/501. Montreal, Canada: ICAO.
International Civil Aviation Organization (ICAO) . (2004). Manual on the Implementation of ICAO Language Proficiency Requirements Document 9835AN453. Montreal, Canada: ICAO.
International Civil Aviation Organization (ICAO) . (2007). Working Paper Assembly 36th Session A36-WP/151; TE/36.
International Civil Aviation Organization (ICAO) . (2013). Retrieved from www.icao.int/safety/lpr/Pages/Language-Proficiency-Requirements.aspx.
Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? Language Testing 35 405–426. https://doi.org/10.1177/0265532218758127.
Kim, H. and Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. Australian Review of Applied Linguistics 32: 23.1–17. https://doi.org/10.2104/aral0923.
Kim, H. and Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test-takers in Korea. Language Testing 32 129–149. https://doi.org/10.1177/0265532214544394.
Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. Language in Society 17 375–399. Retrieved from www.jstor.org/stable/4167952.
Mathews, E. (2004). New provisions for English language proficiency are expected to improve aviation safety. ICAO Journal 59: 4–6.
McNamara, T. (1996). Measuring Second Language Performance. London, UK: Longman.
McNamara, T. (2014). Epilogue: Thirty years on: Evolution or revolution? Language Assessment Quarterly 11 226–232. https://doi.org/10.1080/15434303.2014.895830.
McNamara, T. and Roever, C. (2006). Language Testing: The Social Dimension. Malden, MA: Wiley-Blackwell.
Mell, J. (2004). Language training and testing in aviation needs to focus on job-specific competencies. ICAO Journal 59: 12–14.
Moder, C. L. and Halleck, G. B. (2009). Planes, politics and oral proficiency: Testing international air traffic controllers. Australian Review of Applied Linguistics 32: 25.1–25.16. https://doi.org/10.2104/aral0925.
Moder, C. L. and Halleck, G. B. (2010). Can We Get a Little Higher? Proficiency Levels in Aviation English. Presentation at the Language Testing Research Colloquium, Cambridge, UK.
Nevile, M. (2004). Beyond the Black Box: Talk-in Interaction in the Airline Cockpit. Aldershot, UK: Ashgate.
O'Sullivan, B. (2012). Assessment issues in Languages for specific purposes. The Modern Language Journal 96 71–88. https://doi.org/10.1111/j.1540-4781.2012.01298.x.
Read, J. and Knoch, U. (2009). Clearing the air: Applied linguistic perspectives on aviation communication. Australian Review of Applied Linguistics 32: 21.1–21.11. https://doi.org/10.2104/aral0921.
Sänne, J. M. (1999). Creating Safety in Air Traffic Control. Lund, Sweden: Arkiv Forlag.
Sinclair, J. (1991). Corpus Concordance Collocation. Oxford, UK: Oxford University Press.
Teasdale, A. (1996). Content validity in tests for well-defined LSP domains: An approach to defining what is to be tested. In M. Milanovich and N. Saville (eds.), Performance Testing Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium Cambridge and Arnhem. Cambridge, UK/: Cambridge University Press, 211–230.
Van Moere, A. , Suzuki, M. , Downey, R. and Cheng, J. (2009). Implementing ICAO language proficiency requirements in the versant aviation English test. Australian Review of Applied Linguistics 32: 27.1–27.17. https://doi.org/10.2104/aral0927.
Wyss-Bühlmann, E. (2005). Variation and Co-operative Communication Strategies in Air Traffic Control English. Bern, Switzerland: Peter Lang.

Revisiting language assessment for immigration and citizenship

Canagarajah, S. (ed.). (2017). The Routledge Handbook of Migration and Language. New York, NY: Routledge. This collection of 31 chapters is an exploration of concepts, contexts, methods, and policies related to immigration, citizenship, and language.
Cisneros, J. D. (ed.). (2013). The Border Crossed U.S. Tuscaloosa, AL: University of Alabama Press. This collection of essays offers new ways of thinking of the border between the US and Mexico by examining race, coloniality and citizenship, and radical and hybrid rhetoric social movements related to immigration and citizenship.
Hartelius, E. Johanna . (2015). The Rhetorics of U.S. Immigration: Identity, Community, Otherness. State College, PA: Penn State University Press. This monograph examines US immigration as a rhetorical process

that invents “persons and communities with respect to space and place” (p. 1).
Americans with Disabilities Act of 1990, Pub. L. No. 101–336, § 2, 104 Stat. 328. (1991).
Crawford, J. (1992). Language loyalties. Chicago, IL: University of Chicago Press.
De Valle, S. (2003). Language Rights and the Law. Bristol, UK: Multilingual Matters.
Etzioni, A. (2007). Citizenship tests: A comparative, communitarian perspective. Political Quarterly 78 353–363. https://doi.org/10.1111/j.1467-923X.2007.00864.x.
Gambino, C. , Acosta, Y. and Grieco, E. (2014). English-Speaking Ability of the Foreign-Born Population in the United States. American Community Survey Reports 26. Washington, DC: US Census Bureau.
Garcia, J. (2010). Is the U.S. Naturalization Test in Violation of the 14th Amendment of the U.S. Constitution? Paper presented at the SCALAR 13 Conference, UCLA, Los Angeles.
Huntington, S. (2005). Who Are We? America’s Great Debate. New York, NY: Free Press.
The Immigration Act of 1924. United States. Congress. Senate. Committee on Immigration. (1928). Restriction of western hemisphere immigration, etc. February 1, 27–29; March 1 and 5, 1928. US Government Printing Office.
The Immigration and Nationality Act of 1952. Pub. L. No. 82-414, Stat. 163, Title 8 of the U.S. Code, Chapter 12. USCIS, Department of Homeland Services.
Jones, M. (2018). Birthright Citizens. Cambrige, UK: CambridgeUniversity Press.
Kennedy, J. F. (1964). A Nation of Immigrants. New York, NY: Harper.
Kunnan, A. J. (2009a). Politics and legislation in immigration and citizenship testing: The U.S. case. Annual Review of Applied Linguistics 29 37–48. https://doi.org/10.1017/S0267190509090047.
Kunnan, A. J. (2009b). The U.S. Naturalization Test. Language Assessment Quarterly 6 89–97. https://doi.org/10.1080/15434300802606630.
Kunnan, A. J. (2013). Language assessment for immigration and citizenship. In G. Fulcher and F. Davidson (eds.), The Handbook of Language Testing. New York, NY/: Routledge, 152–166.
Kunnan, A. J. (2018). Evaluating Language Assessments. New York, NY: Routledge.
Kunnan, A. J. , Yao, D. and Yang, F. (2019). Language Requirements for Immigration and Citizenship: A Database. Unpublished manuscript, University of Macau.
Kymlicka, W. (1995). Multicultural Citizenship: A Liberal Theory of Minority Rights. Oxford, UK: Clarendon Press.
Lee, E. (2019). America for Americans: A History of Xenophobia in the United States. New York, NY: Basic Books.
Lew-Williams, B. (2018). The Chinese Must Go. Harvard, MA: Harvard University Press.
Martinez, J. (2010). Why Are Eligible Latinos Not Becoming U.S. Citizens? Seminar paper, TESL 567b, California State University, Los Angeles.
May, S. (2005). Language rights: Moving the debate forward. Journal of Sociolinguistics 9 319–347. https://doi.org/10.1111/j.1360-6441.2005.00295.x.
May, S. (2008). Language education, pluralism, and citizenship. In S. May and N. Hornberger (eds.), Encyclopedia of Language and Education. Language Policy and Political Issues in Education, 2nd edn., vol. 1. Amsterdam, The Netherlands: Springer Science, 15–29.
Min, K. (2010). Is the U.S. Naturalization Test Meaningful in Achieving its Purposes of ‘Civic Nationalism,’ ‘Social Integration,’ and ‘Political Allegiance’ with Citizens of Korean Origin? Seminar paper, TESL 567b, California State University, Los Angeles.
Ngai, M. (2004). Impossible Subjects. Princeton, NJ: Princeton University Press.
The Naturalization Act of 1790. A Bill to Establish a Uniform Rule of Naturalization, and Enable Aliens to Hold Lands under Certain Conditions; 3/4/1790, etc. U.S. Senate, Record Group 46; National Archives Building, Washington, DC.
Parker, K. (2015). Making Foreigners. Cambridge, UK: Cambridge University Press.
Pavlenko, A. (2002). ‘We have room for but one language here’: Language and national identity in the U.S. at the turn of the 20th century. Multilingua 21 163–196. https://doi.org/10.1515/mult.2002.008.
Pickus, N. (2005). True Faith and Allegiance: Immigration and American Civic Nationalism. Princeton, NJ: Princeton University Press.
Pogge, T. (2003). Accommodation rights for Hispanics. In W. Kymlicka and A. Patten (eds.), Language Rights and Political Theory. Oxford, UK/: Oxford University Press, 105–122.
Ricento, T. and Wright, W. (2008). Language policy and education in the United States. In S. May and N. Hornberger (eds.), Encyclopedia of Language Education, 2nd edn., vol. 1. Amsterdam, The Netherlands: Springer, 285–300.
Schlesinger, A. (1992). The Disuniting if America: Reflections on a Multicultural Society. New York, NY: WW Norton.
Thompson, F. (1920). Schooling of the immigrant. [Reprinted 2015]. Andesite Press.
U.S. Citizenship and Immigration Services . (2020). Retrieved December 10, 2020, from http://www.uscis.gov.

Citizenship Resource Center . Retrieved from www.uscis.gov/citizenship.
Scoring guidelines of the test . Retrieved from www.U.S.cis.gov/sites/default/files/U.S.CIS/Office%20of%20Citizenship/Citizenship%20Resource%20Center%20Site/Publications/PDFs/Test_Scoring_Guidelines.pdf.
Civics (history and government) questions for the naturalization test . Retrieved from www.U.S.cis.gov/sites/default/files/document/questions-and-answers/100q.pdf.
Dred Scott v. Sandford, 60 U.S. 393 (1857).
Farrington v. Tokushige, 273 U.S. 284 (1927).
Lau v. Nichols, 414 U.S. 563 (1974).
Meyer v. Nebraska, 262 U.S. 390 (1923).

Classroom-based assessment

Cheng, L. and Fox, J. (2017). Assessment in the Language Classroom. New York, NY: Palgrave Macmillan. This practical resource for language teachers provides an overview of assessment practices which support learning. Informed by empirical research, it combines detailed explanations with hands-on examples from classroom assessment experience.
Hartwick, P. and Nowlan, N. S. (2018). Integrating virtual spaces: Connecting affordances of 3D virtual learning environments to design for twenty-first century learning. In Y. Qian (ed.), Integrating Multi-User Virtual Environments in Modern Classrooms. Hershey, PA: ICI Global, 111–136. Hartwick and Nowlan examine the importance of social interaction, task design, and teacher feedback, given the affordances of online learning.
Moss, P. A. (2016). Shifting the focus of validity for test use. Assessment in Education: Principles, Policy & Practice 23: 236–251. https://doi.org.10.1080/0969594X.2015.1072085. This landmark article proposes a long-term, transdisciplinary research agenda “to support routine conceptual uses” (p. 248) of external test data and suggests strategies for using such “data well” (p. 248), through evidence-rich, empirical partnerships that connect testers with test users (e.g., teachers, administrators, policy makers).
Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin, Germany and Boston, MA: DeGruyter Mouton. Turner and Purpura propose a multi-dimensional framework for learning oriented assessment (LOA) that prioritizes learning when considering the interrelationships across instruction, assessment, and learning.
Abdulhamid, N. (2018). What Is the Relationship Between Alignment and Washback? A Mixed-Methods Study of the Libyan EFL Context. Unpublished doctoral dissertation, Carleton University Ottawa, Canada.
Abdulhamid, N. and Fox, J. (2020). Portfolio based language assessment (PBLA) in language instruction for newcomers to Canada (LINC) programs: Taking stock of teachers’ experience. Canadian Journal of Applied Linguistics 23 168–192. https://doi.org/10.37213/cjal.2020.31121.
Alderson, J. C. (2005). Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment. London, UK: Continuum.
Alderson, J. C. , Brunfaut, T. and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. Applied Linguistics 36 236–260. https://doi.org/10.1093/applin/amt046.
Andrade, H. L. and Brookhart, S. (2019). Classroom assessment as the co-regulation of learning. Assessment in Education: Principles, Policy & Practice 27 350–372. https://doi.org/10.1080/0969594X.2019.1571992.
Andrade, H. L. and Cizek, G. J. (2010). Handbook of Formative Assessment. New York, NY: Routledge.
Artemeva, N. and Fox, J. (2010). Awareness vs. production: Probing students’ antecedent genre knowledge. Journal of Business and Technical Communication 24 476–515. https://doi.org/10.1177%2F1050651910371302.
Assessment Reform Group . (2002). Assessment for Learning: 10 Principles. London, UK: Assessment Reform Group. Retrieved from www.aaia.org.uk/storage/medialibrary/o_1d8j89n3u1n0u17u91fdd1m4418fh8.pdf.
Barrett, H. C. (2007). Researching electronic portfolios and learner engagement: The REFLECT initiative. Journal of Adolescent & Adult Literacy 50 436–449. https://doi:10.1598/JAAL.50.6.2.
Biggs, J. and Tang, C. (2011). Teaching for Quality Learning. Berkshire, UK: McGraw Hill.
Birenbaum, M. , DeLuca, C. , Earl, L. , Heritage, M. , Klenowski, V. , Looney, A. and Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. Policy Futures in Education 13 117–140. https://doi.org/10.1177/1478210314566733.
Black, P. and Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan 92 81–90. https://doi.org/10.1177/003172171009200119.

Blazer, C. (2011). Unintended Consequences of High-Stakes Testing: Information Capsule, Volume 1008: Research Report for the Miami-Dade County Public Schools Research Services. Retrieved from https://archive.org/details/ERIC_ED536512.
Bloom, B. S. (ed.). (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. New York, NY: McKay.
Breen, M. (1987a). Contemporary paradigms in syllabus design: Part I. Language Teaching 20 81–92. https://doi.org/10.1017/S0261444800004365.
Breen, M. (1987b). Contemporary paradigms in syllabus design: Part II. Language Teaching 20 157–174. https://doi.org/10.1017/S026144480000450X.
Bygate, M. , Skehan, P. and Swain, M. (eds.). (2001). Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing. Harlow, UK: Longman.
Callies, M. and Götz, S. (eds.). (2015). Learner Corpora in Language Testing and Assessment. Amsterdam, The Netherlands: John Benjamins.
Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. Language Testing 33 453–472. https://doi.org/10.1177/0265532215593312.
Chapelle, C. A. and Sauro, S. (eds.). (2017). The Handbook of Technology and Second Language Teaching and Learning. Hoboken, NJ: John Wiley & Sons, Inc.
Chatterji, M. (ed.). (2013). Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity. Bingley, UK: Emerald Group.
Cheng, L. and Fox, J. (2017). Assessment in the Language Classroom. New York, NY: Palgrave Macmillan.
Cheng, L. , Watanabe, Y. and Curtis, A. (2004). Washback in Language Testing: Research Contexts and Methods. Mahwah, NJ: Lawrence Erlbaum Associates.
Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics and challenges. In H. L. Andrade and G. J. Cizek (eds.), Handbook of Formative Assessment. New York, NY/: Routledge, 3–17.
Curtis, A. (2017). Methods and Methodologies for Language Teaching. London, UK: Palgrave Macmillan.
Doe, C. (2015). Student interpretations of diagnostic feedback. Language Assessment Quarterly 12 110–135. https://doi.org/10.1080/15434303.2014.1002925.
Earl, L. (2013). Assessment as Learning: Using Classroom Assessment to Maximize Student Learning, 2nd edn. E-reader version. Thousand Oaks, CA: Corwin Press.
Elbow, P. and Belanoff, P. (1986). Portfolios as a substitute for proficiency examinations. College Composition and Communication 37 336–339. Retrieved from www.jstor.org/stable/358050.
Eynon, B. and Gambino, L. (2018). Catalyst in Action: Case Studies of High Impact EPortfolio Practice. Sterling, VA: Stylus Publishing.
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. https://doi.org/10.1016/j.jeap.2008.12.004.
Fox, J. (2014). Portfolio based language assessment (PBLA) in Canadian immigrant language training: Have we got it wrong? Contact 40 68–83. Retrieved from www.teslontario.org/uploads/publications/researchsymposium/ResearchSymposium2014.pdf.
Fox, J. and Artemeva, N. (2017). From diagnosis toward academic support: Developing a disciplinary, ESP-based writing task and rubric to identify the needs of entering undergraduate engineering students. ESP Today 5 148–171. https://doi.org/10.18485/esptoday.2017.5.2.2.
Fox, J. (forthcoming). Reconsidering Context in Language Assessment: Transdisciplinary Perspectives, Social Theories, and Validity. New York, NY: Routledge.
Fox, J. and Hartwick, P. (2011). Taking a diagnostic turn: Reinventing the portfolio in EAP classrooms. In D. Tsagari and I. Csépes (eds.), Classroom-Based Language Assessment. Frankfurt, DE/: Peter Lang, 47–62.
Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. Language Testing 13 208–238. https://doi.org/10.1177%2F026553229601300205.
Gipps, C. V. (1994). Beyond Testing: Towards a Theory of Educational Assessment. London: The Falmer Press.
Hamp-Lyons, L. and Condon, W. (2000). Assessing the Portfolio: Principles for Practice, Theory, and Research. Cresskill: Hampton.
Harding, L. , Alderson, J. C. and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. Language Testing 32 317–336. https://doi.org/10.1177%2F0265532214564505.
Hargreaves, A. , Earl, L. and Schmidt, M. (2002). Perspectives on alternative assessment reform. American Educational Research Journal 39 69–95. https://doi.org/10.3102%2F00028312039001069.
Harlen, W. and Winter, J. (2004). The development of assessment for learning: Learning from the case of science and mathematics. Language Testing 21 390–408. https://doi.org/10.1191%2F0265532204lt289oa.
Hartwick, P. (2018). Exploring the Affordances of Online Learning Environments: 3DVLs and ePortfolios in Second Language Learning and Teaching. Unpublished doctoral dissertation, Carleton University, Ottawa,

Canada.
Hartwick, P. and Nowlan, N. S. (2018). Integrating virtual spaces: Connecting affordances of 3D virtual learning environments to design for twenty-first century learning. In Y. Qian (ed.), Integrating Multi-User Virtual Environments in Modern Classrooms. Hershey, PA/: ICI Global, 111–136.
Heitink, M. C. , Van der Kleij, F. M. , Veldkamp, B. P. , Schildkamp, K. and Kippers, W. B. (2016). A systematic review of prerequisite for implementing assessment for learning in classroom practice. Educational Research Review 17 50–62. https://doi.org/10.1016/j.edurev.2015.12.002.
Hill, K. and McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. Language Testing 29 395–420. https://doi.org/10.1177%2F0265532211428317.
Hymes, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmes (eds.), Sociolinguistics: Selected Readings. Harmondsworth, UK/: Penguin, 269–293.
Information and Privacy Commissioner of Ontario . (2019). A Guide to Privacy and Access to Information in Ontario Schools. Retrieved from www.ipc.on.ca/wp-content/uploads/2019/01/guide-to-privacy-access-in-ont-schools.pdf.
Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. Language Testing 26 31–73. https://doi.org/10.1177%2F0265532208097336.
Jang, E. E. and Wagner, M. (2013). Diagnostic feedback in language classroom. In A. Kunnan (ed.), Companion to Language Assessment. Hoboken, NJ/: Wiley-Blackwell, 693–711.
Jang, E. E. , Wagner, M. and Park, G. (2014). Mixed methods research in language testing and assessment. Annual Review of Applied Linguistics 34 123–153. https://doi.org/10.1017/S0267190514000063.
Kempf, A. (2016). The Pedagogy of Standardized Testing: The Radical Impacts of Educational Standardization in the US and Canada. New York, NY: Palgrave Macmillan.
Lee, Y. W. (2015). Future of diagnostic language assessment. Language Testing 32 295–298. https://doi.org/10.1177%2F0265532214565385.
Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice, and change. Language Assessment Quarterly 1 19–41. https://doi.org/10.1207/s15434311laq0101_3.
Leung, C. and Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. Language Testing 21 335–359. https://doi.org/10.1191%2F0265532204lt287oa.
Little, D. (2005). The common European framework and the European language portfolio: Involving learners and their judgements in the assessment process. Language Testing 22 321–336. https://doi.org/10.1191%2F0265532205lt311oa.
Macqueen, S. , Knoch, U. , Wigglesworth, G. , Nordlinger, R. , Singer, R. , McNamara, T. and Brickle, R. (2019). The impact of national standardized literacy and numeracy testing on children and teaching staff in remote Australian indigenous communities. Language Testing 36 265–287. https://doi.org/10.1177%2F0265532218775758.
McAlpine, M. (2005). E-portfolios and digital identity: Some issues for discussion. E-Learning 2 378–387. https://doi.org/10.2304%2Felea.2005.2.4.378.
Moeller, A. J. , Creswell, W. and Saville, N , (eds,). (2016). Second Language Assessment and Mixed Methods Research. Cambridge, UK: Cambridge University Press.
Moss, P. A. (1994). Can there be validity without reliability? Educational researcher 23 5–12. https://doi.org/10.3102%2F0013189X023002005.
Moss, P. A. (2016). Shifting the focus of validity for test use. Assessment in Education: Principles, Policy & Practice 23 236–251. https://doi.org/10.1080/0969594X.2015.1072085.
Nunan, D. (1988). The Learner-Centred Curriculum. Cambridge, UK: Cambridge University Press.
Panadero, E. , Andrade, H. and Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. Australian Educational Researcher 45 13–31.https://doi.org/10.1007/s13384-018-0258.
Pellegrino, J. W. , Chudowsky, N. and Glaser, R. (eds.). (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academy Press.
Poehner, M. E. and Inbar-Lourie, O. (eds.). (2020). Toward a Reconceptualization of Second Language Classroom Assessment: Praxis and Researcher-Teacher Partnership. Cham, CH: Springer.
Poehner, M. E. and Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). Language Teaching Research 17 323–342. https://doi.org/10.1177/1362168813482935.
Prabhu, R. N. S. (1987). Second Language Pedagogy. Toronto, ON: Oxford University Press.
Purpura, J. E. (2019). Questioning the Currency of Second and Foreign Language Certification Exams. Paper presented at the A&H Distinguished Lecture Series at Teacher College, Columbia University, New York, NY. Retrieved from www.tc.columbia.edu/arts-and-humanities/ah-distinguished-speaker-series/.

Purpura, J. E. and Turner, C. (forthcoming). Learning-Oriented Assessment in Language Classrooms: Using Assessment to Gauge and Promote Language Learning. London, UK: Taylor & Francis Ltd.
Read, J. (ed.). (2016). Post-Admission Language Assessment of University Students. Cham: Springer International.
Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. Language Testing 18 429–462. https://doi.org/10.1111/j.1473-4192.2006.00112.x.
Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: A learning focused interpretation. International Journal of Applied Linguistics 16 164–188. https://doi.org/10.1111/j.1473-4192.2006.00112.x.
Savignon, S. J. (1987). Initiatives in Communicative Language Teaching II: A Book of Readings. Reading, MA: Addison-Wesley.
Savin-Baden, M. (2008). Learning Spaces: Creating Opportunities for Knowledge Creation in Academic Life. New York, NY: McGraw-Hill, Society for Research into Higher Education & Open University Press.
Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler , R. M. Gagne and M. Scriven (eds.), Perspective Son Curricular Evaluation. Chicago, IL/: Rand-McNally, 39–83.
Shohamy, E. (2001). The Power of Tests: A Critical Perspective on the Uses of Language Tests. London, UK: Pearson.
Sultana, N. (2019). An Exploration of Alignment of the Secondary School Certificate (SSC) English Examination with Curriculum and Classroom Instruction: A Washback Study in the Context of Bangladesh. Unpublished doctoral dissertation, Queen's University, Kingston, Canada.
Taylor, L. (2009). Developing assessment literacy. Annual Review of Applied Linguistics 29 21–36. https://doi.org/10.1017/S0267190509090035.
Tishman, S. , Jay, E. and Perkins, D. N. (1993). Teaching thinking dispositions: From transmission to enculturation. Theory into Practice 32 147–153. Retrieved from www.jstor.org/stable/1476695.
Torrance, H. (2015). Blaming the victim: Assessment, examinations, and the responsibilisation of students and teachers in neo-liberal governance. Discourse: Studies in the Cultural Politics of Education 38 83–96. https://doi.org/10.1080/01596306.2015.1104854.
Turner, C. E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. International Journal on Pedagogies and Learning 5 103–123. https://doi.org/10.5172/ijpl.5.1.103.
Turner, C. E. (2012). Classroom assessment. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. New York, NY/: Routledge, 65–78.
Turner, C. E. (2014). Mixed methods research. In A. J. Kunnan (ed.), The Companion to Language Assessment. Chichester, UK: John Wiley & Sons Ltd., 1403–1417. doi:10.1002/9781118411360.wbcla142.
Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin, Germany and Boston, MA: DeGruyter Mouton.
Upshur, J. A. and Turner, C. E. (1995). Constructing rating scales for second language tests. English Language Teaching Journal 49 3–12. https://doi.org/10.1093/elt/49.1.3.
Van den Branden, K. (2016). The role of teachers in task-based language education. Annual Review of Applied Linguistics 36 164–181. doi:10.1017/S0267190515000070.
van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. In J. P. Lantolf (ed.), Sociocultural theory and second language learning. Oxford, UK/: Oxford University Press, 245–259.
Van Viegen Stille, S. , Bethke, R. , Bradley-Brown, J. , Giberson, J. and Hall, G. (2016). Broadening educational practice to include translanguaging: An outcome of educator inquiry into multilingual students' learning needs. Canadian Modern Language Review 72 480–503. https://doi.org/10.3138/cmlr.3432.
Vogt, K. and Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. Language Assessment Quarterly 11 374–402. https://doi.org/10.1080/15434303.2014.960046.
Wiggins, G. and McTighe, J. (2005). Understanding by Design. Alexandria, VA: Association for Supervision and Curriculum Development.
Wiliam, D. (2011). What is assessment for learning? Studies in Educational Evaluation 37 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001.
Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts , P. R. Pintrich and M. Zeidner (eds.), Handbook of Self-Regulation. San Diego, CA: Academic Press, 13–39. doi:10.1016/b978-012109890-2/50031-7.

Washback

Cheng, L. , Sun, Y. and Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. Language Teaching 48: 436–470. doi:10.1017/S0261444815000233 This is one of the most up-to-date systematic reviews of washback studies in the field so far. A systematic search of the pertinent literature between 1993 and 2013 identified a total of 123 publications consisting of 36 review articles and 87 empirical studies. This review focuses on the empirical studies only. A further breakdown of these empirical studies reveals 11 books and monographs, 27 doctoral dissertations, 40 journal articles, and 9 book chapters. This intensity of research activity underscores the timeliness and importance of this research topic and highlights its maturity.

Tsagari, D. and Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. G. Or and S. May (eds.), Language Testing and Assessment, 3rd edn. New York, NY: Springer, 359–372. Tsagari and Cheng elaborated the terms *washback*, *impact*, and *consequences* by presenting the major empirical washback studies conducted over the past 30 years. This is a second edition of this chapter (see Cheng, 2008, as the first edition). The complexity of washback research has been discussed by pointing out the problems and difficulties that washback studies face. The key strength of this chapter lies in providing detailed future directions of washback studies, which are valuable for both novice and experienced researchers.

Abdulhamid, N. (2018). What Is the Relationship Between Alignment and Washback? A Mixed-Methods Study of the Libyan EFL Context. Doctoral dissertation, Carleton University, Ottawa, Canada. Retrieved from https://curve.carleton.ca/2fe0aa90-9afd-498d-8e6d-78f96f055ff4.

Ahmed, A. A. M. (2018). Washback: Examining English Language Teaching and Learning in Libyan Secondary School Education. Doctoral dissertation, The University of Huddersfield, Huddersfield, UK. Retrieved from http://eprints.hud.ac.uk/id/eprint/34799/.

Alderson, J. C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. Language Testing 13 280–297. https://doi.org/10.1177/026553229601300304.

Alderson, J. C. and Wall, D. (1993). Does washback exist? Applied Linguistics 14 115–129. https://doi.org/10.1093/applin/14.2.115.

Anand, P. (2018). Testing Regime Change as Innovation: Washback Potential Over Time. Doctoral dissertation, Carleton University, Ottawa, Canada. Retrieved from https://curve.carleton.ca/system/files/etd/22e85455-2126-4496-a5f2-1b17277a5cb1/etd_pdf/48581ca9698e9159a3f830896ba92cc0/anand-testingregimechangeasinnovationwashbackpotential.pdf.

Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quarterly 2 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests, vol. 1. New York, NY: Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. Language Testing 13 257–279. https://doi.org/10.1177/026553229601300303.

Bailey, K. M. (1999). Washback in Language Testing. Princeton, NJ: Educational Testing Service.

Bhattarai, Y. B. (2019). The Assessment System Reform of the School Leaving Certificate Exam in Nepal: A Grounded Theory of the Reform Process. Doctoral dissertation, Ottawa, Ontario, Canada. Retrieved from https://ruor.uottawa.ca/handle/10393/39002.

Biggs, J. (1995). Assumptions underlying new approaches to educational assessment: Implications for Hong Kong. Curriculum Forum 4: 1–22.

Booth, D. K. (2018). The sociocultural activity of high stakes standardised language testing, vol. 12. Cham: Springer. https://doi.org/10.1007/978-3-319-70446-3_4.

Booth, D. K. (2012). Exploring the Washback of the TOEIC in South Korea: A Sociocultural Perspective on Student Test Activity. Doctoral dissertation, The University of Auckland, Auckland, New Zealand. Retrieved from http://researchspace.auckland.ac.nz.

Brunfaut, T. (2014). A lifetime of language testing: An interview with J. Charles Alderson. Language Assessment Quarterly 11 103–119. https://doi.org/10.1080/15434303.2013.869818.

Chapelle, C. A. (2020). An introduction to language testing's first virtual special issue: Investigating consequences of language test use. Language Testing 37 638–645. https://doi.org/10.1177/0265532220928533.

Chen, L. (2002). Taiwanese Junior High School English Teachers' Perceptions of the Washback Effect of the Basic Competence Test in English. Doctoral dissertation, Ohio State University, Columbus, OH. Retrieved from https://etd.ohiolink.edu/.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. Language and Education 11 38–54. https://doi.org/10.1080/09500789708666717.

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. Studies in Educational Evaluation 24 279–301. https://doi.org/10.1016/S0191-

491X(98)00018-2.
Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. Teaching and Teacher Education 15 253–271. https://doi.org/10.1016/S0742-051X(98)00046-8.
Cheng, L. (2005). Changing Language Teaching Through Language Testing: A Washback Study, vol. 21. Cambridge, UK: Cambridge University Press.
Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy and N. H. Hornberger (eds.), Encyclopedia of Language and Education. New York, NY/: Springer, 2479–2494.
Cheng, L. (2014). Consequences, impact, and washback. In A. J. Kunnan (ed.), The Companion to Language Assessment. Hoboken, NJ/: John Wiley & Sons, 1–14.
Cheng, L. (2018). Geopolitics of assessment. In S. Abrar-ul-Hassan (ed.), TESOL Encyclopedia of English Language Teaching: Teaching English as an International Language. Hoboken, NJ: John Wiley & Sons.
Cheng, L. , Sun, Y. and Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. Language Teaching 48 436–470. https://doi.org/10.1017/S0261444815000233.
Cheng, L. , Watanabe, Y. and Curtis, A. (eds.). (2004). Washback in Language Testing: Research Contexts and Methods. Mahwah, NJ: Lawrence Erlbaum.
Crocker, L. (2006). Preparing examinees for test taking: Guidelines for test developers and test users. In S. Downing and T. Haladyna (eds.), Handbook of Test Development. Mahwah, NJ/: Lawrence Erlbaum, 115–128.
Cronbach, L. (1971). Test validation. In R. L. Thorndike (ed.), Educational Measurement, 2nd edn. Washington, DC: American Council on Education, 443–507.
Fox, J. and Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario secondary school literacy test by first and second language test-takers. Assessment in Education 14 9–26. https://doi.org/10.1080/09695940701272773.
Froehlich, V. (2016). Washback of an oral exam on teaching and learning in German middle schools. In D. Tsagari (ed.), Classroom-Based Assessment in L2 Contexts. Newcastle upon Tyne, UK/: Cambridge Scholars Publishing, 161–185.
Froetscher, D. (2016). A new national exam: A case of washback. In D. Tsagari and J. Banerjee (eds.), Contemporary Second Language Assessment. Oxford/: Bloomsbury Academic, 61–81.
Fullan, M. (2015). The New Meaning of Educational Change. New York, NY: Routledge.
Green, A. (2007). IELTS Washback in Context: Preparation for Academic Writing in Higher Education. Cambridge, UK: Cambridge University Press.
Henrichsen, L. E. (1989). Diffusion of Innovations in English Language Teaching: The ELEC Effort in Japan, 1956–1968. New York, NY: Greenwood Press.
Hoque, M. E. (2011). Washback of the Public Examination on Teaching and Learning English as a Foreign Language (eFL) at the Higher Secondary Level in Bangladesh. Unpublished doctoral dissertation. Jahangirnagar University, Dhaka, Bangladesh.
Hughes, A. (1993). Backwash and TOEFL 2000. University of Reading, England, Unpublished manuscript.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement 50 1–73. https://doi.org/10.1111/jedm.12000.
Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), Educational Measurement, 4th edn. Washington, DC: American Council on Education, Praeger, 17–64.
Khaniya, T. R. (1990). Examinations as Instruments for Educational Change: Investigating the Washback Effect of the Nepalese English Exams. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK.
Kim, E. Y. J. (2017). The TOEFL iBT writing Korean students' perceptions of the TOEFL iBT writing test. Assessing Writing 33 1–11. https://doi.org/10.1016/j.asw.2017.02.001.
Koh, K. , Burke, L. E. C. A., Luke, A. , Gong, W. and Tan, C. (2018). Developing the assessment literacy of teachers in Chinese language classrooms: A focus on assessment task design. Language Teaching Research 22 264–288. https://doi.org/10.1177/1362168816684366.
Kunnan, A. J. (2009). Politics and legislation in citizenship testing in the U.S. Annual Review of Applied Linguistics 29 37–48. https://doi.org/10.1017/S0267190509090047.
Latham, H. (1877). On the Action of Examinations Considered as a Means of Selection. Cambridge, UK: Deighton, Bell and Company.
Ma, J. (2019). Did test preparation practices for the college English test work? A study from Chinese students' perspective. In S. Papageorgiou and K. Bailey (eds.), Global Perspective on Language Assessment. New York, NY/: Routledge, 169–182.
Ma, J. and Cheng, L. (2016). Chinese student' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth and significance. TESL Canada Journal 33 58–79. https://doi.org/10.18806/tesl.v33i1.1227.
Markee, N. (1997). Managing Curricular Innovation. Cambridge: Cambridge University Press.

Marshall, B. (2017). The politics of testing. English in Education 51 27–43. https://doi.org/10.1111/eie.12110.
Mathew, R. (2012). Understanding washback: A case study of a new exam in India. In C. Tribble (ed.), Managing Change in English Language Teaching: Lessons from Experience. London/: British Council, 195–202.
McNamara, T. (2010). The use of language tests in the service of policy: Issues of validity. Revue française de linguistique appliquée xv 7–23. https://doi.org/10.3917/rfla.151.0007.
Memon, N. (2015). 'If I Just Get One IELTS Certificate, I Can Get Anything': An Impact Study of IELTS in Pakistan. Doctoral dissertation, The University of Edinburgh, The University of Edinburgh. Retrieved from https://era.ed.ac.uk/handle/1842/21122.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Ace and Macmillan, 13–103.
Messick, S. (1996). Validity and washback in language testing. Language Testing 13 241–256. https://doi.org/10.1177/026553229601300302.
Onaiba, A. M. E. (2013). Investigating the Washback Effect of a Revised EFL Public Examination on Teachers' Instructional Practices, Materials and Curriculum. Doctoral dissertation, University of Leicester, UK. Retrieved from https://pdfs.semanticscholar.org/8e6c/c63f0213d4f965e1e9b19dd0b8d39d5bd3f7.pdf.
Oo, C. Z. (2019). Assessment for Learning Literacy and Pre-Service Teacher Education: Perspectives from Myanmar. Unpublished doctoral dissertation. The University of New South Wales, Australia.
Pan, Y. C. (2016). Learners' perspectives of factors influencing gains in standardized English test scores. TEFLIN Journal 27 63–81. https://doi.org/10.15639/teflinjournal.v27i1/63-81.
Pan, Y. C. and Newfields, T. (2013). Student washback from tertiary standardized English proficiency exit requirements in Taiwan. Journal of Teaching and Learning 9 1–16. https://doi.org/10.22329/JTL.V9I1.3540.
Pellegrino, J. W. , DiBello, L. V. and Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. Educational Psychologist 51 59–81. https://doi.org/10.1080/00461520.2016.1145550.
Popham, W. J. (1987). The merits of measurement-driven instruction. Phi Delta Kappan 68: 679–682.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. Language Testing 22 142–173. https://doi.org/10.1191/0265532205lt300oa.
Rind, I. A. and Mari, M. A. (2019). Analysing the impact of external examination on teaching and learning of English at the secondary level education. Cogent Education 6 1–14. https://doi.org/10.1080/2331186X.2019.1574947.
Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behavior during test preparation. Papers in Language Testing and Assessment 8: 69–95.
Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? Educational Measurement 9 15–22. https://doi.org/10.1111/j.1745-3992.1990.tb00374.x.
Shih, C. (2007). A new washback model of students' learning. The Canadian Modern Language Review 64 135–162. https://doi.org/10.3138/cmlr.64.1.135.
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? Language Testing 14 340–349. https://doi.org/10.1177/026553229701400310.
Shohamy, E. (1998). Critical language testing and beyond. Studies in Educational Evaluation 24 331–345. https://doi.org/10.1016/S0191-491X(98)00020-0.
Shohamy, E. (2001). Democratic assessment as an alternative. Language Testing 18 373–391. https://doi.org/10.1177/026553220101800404.
Shohamy, E. , Donitsa-Schmidt, S. and Ferman, I. (1996). Test impact revisited: Washback effect over time. Language Testing 13 298–317. https://doi.org/10.1177/026553229601300305.
Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. Language Teaching Research 9 5–29. https://doi.org/10.1191/1362168805lr152oa.
Sultana, N. (2018). A brief review of washback studies in the South Asian countries. The Educational Review, USA 2 468–474. https://doi.org/10.26855/er.2018.09.002.
Sultana, N. (2019a). Exploring the Alignment of the Secondary School Certificate English Examination with Curriculum and Classroom Instruction: A Washback Study in Bangladesh. Doctoral dissertation, Queen's University, Kingston, Ontario, Canada. Retrieved from https://qspace.library.queensu.ca/handle/1974/26482.
Sultana, N. (2019b). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh. Language Testing in Asia 9 1–14. https://doi.org/10.1186/s40468-019-0077-8.
Sun, Y. (2016). Context, Construct, and Consequences: Washback of the College English Test in China. Doctoral dissertation, Queen's University, Kingston, Canada. Retrieved from https://qspace.library.queensu.ca/handle/1974/13985.
Tan, M. and Turner, C. E. (2015). The impact of communication and collaboration between test developers and teachers on a high-stakes ESL exam: Aligning external assessment and classroom practices. Language Assessment Quarterly 12 29–49. https://doi.org/10.1080/15434303.2014.1003301.

Tsagari, D. and Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy , I. G. Or and S. May (eds.), Language Testing and Assessment, 3rd edn. New York, NY: Springer, 359–372.
Umashankar, S. (2017). Washback Effects of Speaking Assessment of Teaching English in Sri Lankan Schools. Doctoral dissertation, University of Bedfordshire, Luton, England. Retrieved from https://core.ac.uk/download/pdf/151440224.pdf.
Vogt, K. , Tsagari, D. and Spanoudis, G. (2020). What do teachers think they want? A comparative study of in-service language teachers' beliefs on LAL training needs. Language Assessment Quarterly 17 386–409. https://doi.org/10.1080/15434303.2020.1781128.
Volante, L. , DeLuca, C. , Adie, L. , Baker, E. , Harju-Luukkainen, H. , Heritage, M. , Schneider, C. , Stobart, G. , Tan, K. and Wyatt-Smith, C. (2020). Synergy and tension between large-scale and classroom assessment: International trends. Educational Measurement: Issues and Practice 39 21–29. https://doi.org/10.1111/emip.12382.
Vyn, R. (2019). Promoting Curricular Innovation Through Language Performance Assessment: Leveraging AAPPL Washback in a K–12 World Languages Program. Doctoral dissertation, University of Iowa, Iowa City, Iowa, the USA. Retrieved from https://ir.uiowa.edu/etd/6872/.
Wall, D. (2005). The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory, vol. 22. Cambridge, UK: Cambridge University.
Wall, D. (2012). Washback. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London, UK/: Routledge, 79–92.
Wall, D. and Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. Language Testing 10 41–69. https://doi.org/10.1177/026553229301000103.
Wall, D. and Horák, T. (2007). Using baseline studies in the investigation of test impact. Assessment in Education 14 99–116. https://doi.org/10.1080/09695940701272922.
Wang, S. (2018). Investigating the Consequential Validity of the Hanyu Shuiping Kaoshi (Chinese Proficiency Test) by Using an Argument-Based Framework. Doctoral dissertation, McGill University, Montreal, Quebec, Canada. Retrieved from https://escholarship.mcgill.ca/concern/theses/4q77ft93g.
Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. Language Testing 13 318–333. https://doi.org/10.1177/026553229601300306.
Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. NISE Research Monograph No. 6. Madison: University of Wisconsin-Madison, National Institute for Science Education.
Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. Applied Measurement in Education 20 7–25. https://doi.org/10.1080/08957340709336728.
Weir, C. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke: Palgrave Macmillan.
Xie, Q. (2010). Test Design and Use, Preparation, and Performance: A Structural Equation Modeling Study of Consequential Validity. Unpublished doctoral dissertation. University of Hong Kong, Hong Kong.

Assessing young learners

Nikolov, M. (ed.). (2016). Assessing Young learners of English: Global and Local Perspectives. New York, NY: Springer.
Wolf, M. and Butler, Y. G. (eds.). (2017). English Language Proficiency for Young Learners. New York, NY: Routledge.
Prošić-Santovac, D. and Rixon, S. (eds.). (2019). Integrating Assessment into Early Language Learning and Teaching. Bristol, UK: Multilingual Matters. The three edited books gather recent research on assessment for YLLs, but each book has a slightly different focus. Nikolov (2016) considers different assessment practices around the world and discusses the unique challenges associated with assessing YLLs. Wolf and Butler (2017) offer theoretical and empirical accounts for developing and validating assessment for YLLs, mainly focusing on large-scale assessments in English. Prošić-Santovac and Rixon (2019) turn their attention to integrating assessment with pedagogy in the context of foreign language education in Europe.
Papp, S. and Rixon, S. (eds.). (2018). Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners. Cambridge, UK: Cambridge University Press. As part of the Studies in Language Testing Series of Cambridge English Assessment, this volume focuses on YLLs (ages 6–16) and surveys how language assessment constructs for YLLs have been conceptualized and operationalized in language tests. This is a highly comprehensive and readable book, particularly for testing development professionals.

Nikolov, M. and Timpe-Laughlin, V. (2020). Assessing young learners' foreign language abilities. Language Teaching. https://doi.org/10.1017/s0261444820000294. This is a state-of-the-art review article on assessment of young learners that focuses on foreign language contexts. The review offers a critical and yet accessible review on the major issues concerning assessments for young foreign language learners. The paper concludes with ten questions for future directions, which can be an excellent guide for both researchers and practitioners.

Abedi, J. , Hofstetter, C. H. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. Review of Educational Research 74 1–28. https://doi.org/10.3102/00346543074001001.

Alexiou, T. , Roghani, S. and Milton, J. (2019). Assessing the vocabulary knowledge of preschool language learners. In D. Prošić-Santovac and S. Rixon (eds.), Integrating Assessment into Early Language Learning and Teaching. Bristol, UK/: Multilingual Matters, 207–220.

Artiles, A. J. , Rueda, R. , Salazar, J. J. and Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. Exceptional Children 71 283–300. https://doi.org/10.1177/001440290507100305.

Bailey, A. L. (2017). Progressions of a new language: Characterizing explanation development for assessment with young language learners. Annual Review of Applied Linguistics 37 241–263. https://doi.org/10.1017/s0267190517000113.

Ballantyne, K. G. (2013). Disproportional representation of English learners among students identified with disabilities: Equity and the federal accountability system. In D. Tsagari and G. Spanoudis (eds.), Assessing L2 Students with Learning and Other Disabilities. Newcastle upon Tyne, UK/: Cambridge Scholars, 3–25.

Becker, C. (2015). Assessment and portfolios. In J. Bland (ed.), Teaching English to Young Learners: Critical Issues in Language Teaching with 3–12 Year Olds. London, UK/: Bloomsbury, 261–278.

Benigno, V. and de Jong, J. (2016). The "global scale of English learning objectives for young learners": A CEFR-based inventory of descriptors. In M. Nikolov (ed.), Assessing Young Learners of English: Global and Local Perspectives. New York, NY/: Springer, 43–64.

Black, P. J. and Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education 5 7–74. https://doi.org/10.1080/0969595980050102.

Boals, T. , Kenyon, D. M. , Blair, A. , Cranley, M. E. , Wilmes, C. and Wright, L. J. (2015). Transformation in K-12 English language proficiency assessment: Changing contexts, changing constructs. Review of Research in Education 39 122–164. https://doi.org/10.3102/0091732x14556072/.

Brindley, G. (2001). Language assessment and professional development. In C. Elder , A. Brown , K. Hill , N. Iwashita , T. Lumley , T. McNamara and K. O'Loughlin (eds.), Experimenting with Uncertainty: Essays in Honour of Alan Davies. Cambridge, UK/: Cambridge University Press, 126–136.

Burke, A. M. , Morita-Mullaney, T. and Smith, M. (2016). Indiana emerging bilingual student time to reclassification: A survey analysis. American Educational Research Journal 53 1310–1342. https://doi.org/10.3102/0002831216667481.

Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning. In M. Nikolov (ed.), Assessing Young Learners of English: Global and Local Perspectives. New York, NY/: Springer, 291–315.

Butler, Y. G. (2018a). The role of context in young learners' processes for responding to self-assessment items. The Modern Language Journal 102 242–261. https://doi.org/10.1111/modl.12459.

Butler, Y. G. (2018b). Young learners' processes and rationales for responding to self-assessment items: Cases for generic can-do and five-point Likert-type formats. In J. Davis et al. (eds.), Useful Assessment and Evaluation in Language Education. Washington, DC/: Georgetown University Press, 21–39.

Butler, Y. G. (2019). Assessment of young English learners in instructional settings. In A. Gao (ed.), Second Handbook of English Language Teaching. New York, NY/: Springer, 477–495.

Butler, Y. G. and Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. The Modern Language Journal 90 506–518. https://doi.org/10.1111/j.1540-4781.2006.00463.x.

Butler, Y. G. and Lee, J. (2010). The effect of self-assessment among young learners. Language Testing 27 5–31. https://doi.org/10.1177/0265532209346370.

Butler, Y. G. and Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessment. Language Assessment Quarterly 11 45–75. https://doi.org/10.1080/15434303.2013.869814.

Butler, Y. G. and Zeng, W. (2015). Young learners' interactional development in task-based paired-assessment in their first and foreign languages: A case of English learners in China. Education 3–13 43 292–321. https://doi.org/10.1080/03004279.2013.813955.

Carroll, P. E. and Bailey, A. L. (2016). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. Language Testing 33 23–52. https://doi.org/10.1177/0265532215576380.

Chalhoub-Deville, M. B. (2019). Multilingual testing constructs: Theoretical foundations. Language Assessment Quarterly 16 472–480. https://doi.org/10.1080/15434303.2019.1671391.

Chaudron, S. , Di Gioia, R. and Gemo, M. (2017). Young Children (0–8) and Digital Technology: A Qualitative Study Across Europe. Retrieved from https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/young-children-0-8-and-digital-technology-qualitative-study-across-europe.

Choi, I. , Wolf, M. K. , Pooler, E. , Sova, L. and Faulkner-Bond, M. (2019). Investigating the benefits of scaffolding in assessments of young English learners: A case for scaffolding retell tasks. Language Assessment Quarterly 16 161–179. https://doi.org/10.1080/15434303.2019.1619180/.

Cook, V. J. (1992). Evidence for multicompetence. Language Learning 42 557–591. https://doi.org/10.1111/j.1467-1770.1992.tb01044.x.

Council of Europe . (2001/2018). The Common European Framework of Reference for Languages: Learning, Teaching and Assessment. Retrieved from www.coe.int/en/web/common-european-framework-reference-languages/home.

Council of Europe . (2019). ELP Check Lists for Young Learners. Retrieved from www.coe.int/en/web/portfolio/the-language-passport.

Courtney, L. and Graham, S. (2019). “It’s like having a test but in a fun way”: Young learners’ perceptions of a digital game-based assessment on early language learning. Language Teaching for Young Learners 1 161–186. https://doi.org/10.1075/ltyl.18009.cou.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. Working Papers on Bilingualism 19 121–129. Retrieved from https://files.eric.ed.gov/fulltext/ED184334.pdf.

Dann, R. (2002). Promoting Assessment as Learning: Improving the Learning Process. New York, NY: Routledge.

Edelenbos, P. and Kubanek-German, A. (2004). Teacher assessment: The concept of “diagnostic competence.” Language Testing 21 259–283. https://doi.org/10.1191/0265532204lt284oa.

Ellis, G. and Rixon, S. (2019). Assessment for learning with younger learners: Is thinking about their learning a step too far? In D. Prošić-Santovac and S. Rixon (eds.), Integrating Assessment into Early Language Learning and Teaching. Bristol, UK/: Multilingual Matters, 87–104.

Enever, J. (ed.). (2011). ELLiE: Early Language Learning in Europe. London, UK: The British Council.

Enever, J. (2018). Policy and Politics in Global Primary English. Oxford, UK: Oxford University Press.

Firth, J. , Torous, J. , Stubbs, B. , Firth, J. A. , Steiner, G. Z. , … Sarris, J. (2019). The “online brain”: How the Internet may be changing our cognition. World Psychiatry 18 119–129. https://doi.org/10.1002/wps.20617.

Forsyth, C. M. , Luce, C. , Zapata-Rivera, D. , Jackson, G. T. , Evanini, K. and So, Y. (2019). Evaluating English language learners’ conversations: Man vs. machine. Computer Assisted Language Learning 32 398–417. https://doi.org/10.1080/09588221.2018.1517126.

García Mayo, M. P. and Agirre, A. I. (2016). Task repetition and its impact on EFL children’s negotiation of meaning strategies and pair dynamics: An exploratory study. The Language Learning Journal 44 451–466. https://doi.org/10.1080/09571736.2016.1185799.

Giacomo, D. D. , Ranieri, J. and Lacasa, P. (2017). Digital learning as enhanced learning processing? Cognitive evidence for new insights of smart learning. Frontiers in Psychology 8: 1329. https://doi.org/10.3389/fpsyg.2017.01329.

Ghosn, I. K. (2019). Integrating developmentally appropriate assessment with instruction in the young learner classroom. In D. Prošić-Santovac and S. Rixon (eds.), Integrating Assessment into Early Language Learning and Teaching. Bristol, UK/: Multilingual Matters, 52–68.

Goodier, T. and Szabo, T. (2018). Collated Representative Samples of Descriptors of Language Competences Developed for Young Learners (Aged 7–10 and 11–15 Years). Retrieved from www.coe.int/en/web/common-european-framework-reference-languages/bank-of-supplementary-descriptors.

Gu, L. and Hsieh, C. N. (2019). Distinguishing features of young English language learners’ oral performance. Language Assessment Quarterly 16 180–195. https://doi.org/10.1080/15434303.2019.1605518.

Guzman-Orth, D. A. , Lopez, A. A. and Tolentino, F. (2019). Exploring the use of a dual language assessment task to assess young English learners. Language Assessment Quarterly 16 447–463. https://doi.org/10.1080/15434303.2019.1674314.

Harding, L. and Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin, Germany/: De Gruyter, 413–428.

Hsieh, C. N. and Wang, Y. (2019). Speaking proficiency of young language students: A discourse- analytic study. Language Testing 36 27–50. https://doi.org/10.1177/0265532217734240.

Huang, B. and Butler, Y. G. (2020). Validity considerations for assessing language proficiency for young language minority students. Language Assessment Quarterly. https://doi.org/10.1080/15434303.2020.1826486.

Huang, F. L. and Konold, T. R. (2014). A latent variable investigation of the phonological awareness literacy screening-kindergarten assessment: Construct identification and multigroup comparisons between Spanish-speaking English-language learners (ELLs) and non-ELL students. Language Testing 31 205–221.

https://doi.org/10.1177/0265532213496773.
Hung, Y. , Samuelson, B. L. and Chen, S. (2016). Relationship between peer- and self-assessment and teacher assessment of young EFL learners' oral presentations. In M. Nikolov (ed.), Assessing Young Learners of English: Global and Local Perspectives. New York, NY/: Springer, 317–338.
Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle , Y. R. Chung and J. Xu (eds.), Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment. Ames, IA/: Iowa State University, 117–131.
Jang, E. E. (2014). Focus on Assessment. Oxford, UK: Oxford University Press.
Jang, E. E. , Dunlop, M. , Wagner, M. , Kim, Y.-H. and Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. Language Learning 63 400–436. https://doi.org/10.1111/lang.12016.
Jang, E. E. , Vincett, J. M. , van der Boom, E. H. , Lau, C. and Yang, Y. (2017). Considering young learners' characteristics in developing diagnostic assessment intervention. In M. K. Wolf and Y. G. Butler (eds.), English Language Proficiency Assessments for Young Learners. New York, NY/: Routledge, 193–213.
Kangas, S. E. N. (2019). English learners with disabilities: Linguistic development and educational equity in jeopardy. In X. Gao (ed.), Second Handbook of English Language Teaching. New York, NY/: Springer, 919–937.
Keane, L. and Griffin, C. P. (2018). Assessing self-assessment: Can age and prior literacy attainment predict the accuracy of children's self-assessment in literacy? Irish Educational Studies 37 127–147. https://doi.org/10.1080/03323315.2018.1449001.
Kormos, J. (2017). The effects of specific learning difficulties on processes of multilingual language development. Annual Review of Applied Linguistics 37 30–44. https://doi.org/10.1017/s026719051700006x.
Kormos, J. , Brunfaut, T. and Michel, M. (2020). Motivation factors in computer-administered integrated skills tasks: A study of young learners. Language Assessment Quarterly 17 43–59. https://doi.org/10.1080/15434303.2019.1664551.
Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. Journal of Applied Linguistics 1 49–72. https://doi.org/10.1558/japl.1.1.49.55872.
Lee, J. and Butler, Y. G. (2020). Reconceptualizing language assessment literacy: Where are language learners? TESOL Quarterly 54 1098–1111. https://doi.org/10.1002/tesq.576.
Lee, S. and Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. Language Testing 35 239–269. https://doi.org/10.1177/0265532217704009.
Łockiewicz, M. , Sarzała-Przybylska, Z. and Lipowsk, M. (2018). Early predictors of learning a foreign language in pre-school: Polish as a first language, English as a foreign language. Frontiers in Psychology 9: 1813. https://doi.org/10.3389/fpsyg.2018.01813.
McKay, P. (1995). Developing ESL proficiency descriptions for the school context. In G. Brindley (ed.), Language Assessment in Action. Sydney/: National Center for English Language Teaching and Research, 31–63.
McKay, P. (2006). Assessing Young Language Learners. Cambridge, UK: Cambridge University Press.
Menken, K. , Hudson, T. and Leung, C. (2014). Symposium: Language assessment in standards-based education reform. TESOL Quarterly 48 586–614. https://doi.org/10.1002/tesq.180.
Mihaljević Djigunović, J. (2019). Affect and assessment in teaching L2 to young learners. In D. Prošić-Santovac and S. Rixon (eds.), Integrating Assessment into Early Language Learning and Teaching. Bristol, UK/: Multilingual Matters, 19–33.
Mourão, S. and Lourenço, M. (2015). Early Years Second Language Education: International Perspectives on Theory and Practice. New York, NY: Routledge.
Nikolov, M. (2016). Assessing Young Learners of English: Global and Local Perspectives. New York, NY: Springer.
Nikolov, M. and Timpe-Laughlin, V. (2020). Assessing young learners' foreign language abilities. Language Teaching. https://doi.org/10.1017/s0261444820000294.
North, B. (2014). The CEFR in Practice. Cambridge, UK: Cambridge University Press.
Oliver, R. and Azkarai, A. (2019). Patterns of interaction and young ESL learners: What is the impact of proficiency and task type? Language Teaching for Young Learners 1 82–102. https://doi.org/10.1075/ltyl.00006.oli.
Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. Language Testing 6 1–13. https://doi.org/10.1177/026553228900600103.
Papageorgiou, S. and Baron, P. (2017). Using the common European framework of reference to facilitate score interpretations for young learners' English language proficiency assessments. In M. K. Wolf and Y. G. Butler (eds.), English Language Proficiency Assessments for Young Learners. New York, NY/: Routledge, 136–152.
Papp, S. (2018). Test taker characteristics. In S. Papp and S. Rixon (eds.), Examining Young Learners: Research and Practice in Assessing the English of School-Age Children. Cambridge, UK/: Cambridge

University Press, 70–127.
Papp, S. and Rixon, S. (eds.). (2018). Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners. Cambridge, UK: Cambridge University Press.
Papp, S. and 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. New York, NY/: Springer, 139–190.
Pearson Education . (2018). The Global Scale of English Learning Objectives for Young Learners. Harlow, UK: Pearson.
Perry, N. E. and Van de Kamp, K. J. O. (2000). Creating classroom contexts that support young children's development of self-regulated learning. International Journal of Educational Research 33 821–843. https://doi.org/10.1016/s0883-0355(00)00052-5.
Peterson, D. B. and Gillam, R. B. (2013). Predicting reading ability for bilingual Latino children dynamic assessment. Journal of Learning Disabilities 48 3–21. https://doi.org/10.1177/0022219413486930.
Pinter, A. (2015). Task-based learning with children. In J. Bland (ed.), Teaching English to Young Learners: Critical Issues in Language Teaching with 3–12 Year Olds. London, UK/: Bloomsbury, 113–127.
Poehner, M. E. , Zhang, J. and Lu, X. (2017). Computerized dynamic assessment for young language learners. In M. K. Wolf and Y. G. Butler (eds.), English Language Proficiency Assessments for Young Learners. New York, NY/: Routledge, 214–233.
Prošić-Santovac, D. and Navratil, A. (2019). Assessment of very young learners of English and the joy of puppetry: A multiple case study. In D. Prošić-Santovac and S. Rixon (eds.), Integrating Assessment into Early Language Learning and Teaching. Bristol, UK/: Multilingual Matters, 170–187.
Prošić-Santovac, D. and Rixon, S. (eds.). (2019). Integrating Assessment into Early Language Learning and Teaching. Bristol, UK: Multilingual Matters.
Purpura, J. E. (2016). Second and foreign language assessment. The Modern Language Journal 100(supplement) 190–208. https://doi.org/10.1111/modl.12308.
Rea-Dickins, P. (2000). Assessment in early years language learning contexts. Language Testing 17 115–122. https://doi.org/10.1177/026553220001700201.
Rivera, C. and Collum, E. (2006). State Assessment Policy and Practice for English Language Learners: A National Perspective. New York, NY: Routledge.
Rixon, S. (2018). Consequential validity of tests of English for young learners: The impact of assessment on young learners. In S. Papp and S. Rixon (eds.), Examining Young Learners: Research and Practice in Assessing the English of School-Age Children. Cambridge, UK/: Cambridge University Press, 547–587.
Rixon, S. (2019). Developing language curricula for young language learners. In X. Gao (ed.), Second Handbook of English Language Teaching. New York, NY/: Springer, 277–295.
Rixon, S. and Papp, S. (2018). The educational contexts of the teaching of English to young learners and the roles of assessment. In S. Papp and S. Rixon (eds.), Examining Young Learners: Research and Practice in Assessing the English of School-Age Children. Cambridge, UK/: Cambridge University Press, 18–69.
Scarcella, R. (2003). Academic English: A Conceptual Framework (Technical Report No. 2003-1). Irvine, CA: University of California Linguistic Minority Research Institute.
Schissel, J. L. (2019). Social Consequences of Testing for Language-Minoritized Bilinguals in the United States. Clevedon, UK: Multilingual Matters.
Schissel, J. L. , Leung, C. and Chalhoub-Deville, M. (2019). The construct of multilingualism in language testing. Language Assessment Quarterly 16 373–378. https://doi.org/10.1080/15434303.2019.1680679.
Schleppegrell, M. J. (2004). The Language of Schooling: A Functional Linguistics Perspective. New York, NY: Lawrence Erlbaum.
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? Language Testing 14 340–349. https://doi.org/10.1177/026553229701400310.
Sifferlin, A. (2015, April 25). 6-month-old babies are now using tablets and smartphones. TIME. Retrieved October 11, 2019, from https://time.com/3834978/babies-use-devices/.
Smith, A. M. (2013). Developing cognitive assessments for multilingual learners. In D. Tsagari and G. Spanoudis (eds.), Assessing L2 Students with Learning and Other Disabilities. Newcastle upon Tyne, UK/: Cambridge Scholars, 151–167.
Takanishi, R. and Le Menestrel, S. (2017). Promoting the Educational Success of Children and Youth Learning English: Promising Futures. Washington, DC: The National Academic Press.
Teachers of English to Speakers of Other Languages (TESOL) . (1998). Managing the Assessment Process: A Framework for Measuring Student Attainment of the ESL Standards. Alexandria, VA: TESOL.
Téllez, K. and Mosqueda, E. (2015). Developing teachers' knowledge and skills at the intersection of English language learners and language assessment. Review of Research in Education 39 87–121. https://doi.org/10.3102/0091732x14554552.
Timpe-Laughlin, V. and Cho, Y. (2021). Assessing young foreign language learners in global and local contexts. Language Testing (special issue).

Torrance, H. and Pryor, J. (1998). Investigating Formative Assessment. Milton Keynes, UK: Open University Press.
Vogt, P. , de Haas, M. , de Jong, C. , Baxter, P. and Krahmer, E. (2017). Child-robot interactions for second language tutoring to preschool children. Frontiers in Human Neuroscience 11. https://doi.org/10.3389/fnhum.2017.00073.
Vygotsky, L. S. (1978). Mind and Society: The Development of Higher Mental Processes. Cambridge, MA: Harvard University Press.
Winke, P. , Lee, S. , Ahn, J. I. , Choi, I. , Gui, Y. and Yoon, H. (2018). The cognitive validity of child English language tests: What young language learners and their native-speaking peers can reveal. TESOL Quarterly 52 274–303. https://doi.org/10.1002/tesq.396.
Wolf, M. K. and Butler, Y. G. (eds.). (2017). English Language Proficiency for Young Learners. New York, NY: Routledge.
Wolf, M. K. , Guzman-Orth, D. , Lopez, A. , Castellano, K. , Himelfarb, I. and Tsutagawa, F. S. (2016). Integrating scaffolding strategies into technology-enhanced assessments of English learners: Task types and measurement models. Educational Assessment 21 157–175. https://doi.org/10.1080/10627197.2016.1202107.
Wolf, M. K. , Kao, J. C. , Herman, J. , Bachman, L. F.Bailey , A. L. Bachman , P. L. … Chang, S. M. (2008). Issues in Assessing English Language Learners: English Language Proficiency Measures and Accommodation Uses: Literature Review (CRESST Report No. 731). Los Angeles: CRESST.
Wolf, M. K. , Lopez, A. , Oh, S. and Tsutagawa, F. S. (2017). Comparing the performance of young English language learners and native English speakers on speaking assessment tasks. In M. K. Wolf and Y. G. Butler (eds.), English Language Proficiency Assessments for Young Learners. New York, NY/: Routledge, 171–187.
Wong, K. M. and Mak, P. (2019). Self-assessment in the primary L2 writing classroom. The Canadian Modern Language Review/La revue Canadienne des langues vivantes 75 183–196. https://doi.org/10.3138/cmlr.2018-0197.
Yee, N. and Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. Human Communication Research 33 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x.

Dynamic assessment

Lantolf, J. P. and Poehner, M. E. (2011). Dynamic Assessment in the Foreign Language Classroom: A Teacher's Guide, 2nd edn. University Park, PA: Calper. This text is an introduction to the implementation of DA in language contexts. It is aimed at language teachers who wish to bring DA to their own classrooms and provides a brief theoretical background, multiples examples from actual practice, and practical activities for teacher and student development alike.
Lidz, C. (1991). Practitioner's Guide to Dynamic Assessment. New York, NY: Guilford Press. This is a good introduction to practical aspects of interactionist DA. The book summarizes different models of DA in clinical and educational settings, but, most importantly, it provides a step-by-step guide of procedures and tools for conducting DA assessment. The two manuals included in the book provide a good starting point for anyone wishing to start implementing DA.
Poehner, M. (2008). Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting L2 Development. Boston, MA: Springer Science. This is the first comprehensive review of theoretical and practical aspects of DA in second language learning. The first part of this book discusses the origins of DA in Vygotsky's sociocultural theory of mind and subsequent development of different approaches to DA. The second part of the book presents current research in DA applications to L2 settings and offers a model for implementing DA in second language classrooms.
Ableeva, R. (2008). The effects of dynamic assessment on L2 listening comprehension. In J. Lantolf and M. Poehner (eds.), Sociocultural Theory and the Teaching of Second Languages. London, UK: Equinox Pub, 57–86. http://doi.org/10.1558/equinox.29293.
Ableeva, R. (2018). Understanding Learner L2 Development Through Reciprocity. In J. Lantolf , M. Poehner and M. Swain (eds.), The Routledge Handbook of Sociocultural Theory and Second Language Development. London: Routledge, 266–281. https://doi.org/10.4324/9781315624747.
Aljaafreh, A. and Lantolf, J. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. Modern Language Journal 78 465–483. https://doi.org/10.1111/j.1540-4781.1994.tb02064.x.
Andujar, A. (2020). Mobile-mediated dynamic assessment: A new perspective for second language development. ReCALL 1–17. https://doi.org/10.1017/S0958344019000247.
Antón, M. (2009). Dynamic assessment of advanced foreign language learners. Foreign Language Annals 42 576–598. https://doi.org/10.1111/j.1944-9720.2009.01030.x.

Antón, M. (2018). Dynamic diagnosis of second language abilities. In J. Lantolf , M. Poehner and M. Swain (eds.), The Routledge Handbook of Sociocultural Theory and Second Language Development. London, UK: Routledge, 310–323. https://doi.org/10.4324/9781315624747.

Black, P. J. and Williams, D. (1998). Assessment and classroom learning. Assessment in Education 5 7–74. https://doi.org/10.1080/0969595980050102.

Budoff, M. (1987). The validity of learning potential assessment. In C. Lidz (ed.), Dynamic Assessment: An Interactive Approach to Evaluating Learning Potential. New York, NY/: Guilford Press, 52–81.

Campione, J. and Brown, A. (1987). Linking dynamic assessment with school achievement. In C. Lidz (ed.), Dynamic Assessment. New York, NY/: Guilford Press, 82–115.

Davin, K. (2016). Classroom dynamic assessment: A critical examination of constructs and practices. Modern Language Journal 100 813–829. https://doi.org/10.1111/modl.12352.

Davin, K. (2018). Mediator and learner engagement in co-regulated inter-psychological activity. In J. Lantolf , M. Poehner and M. Swain (eds.), The Routledge Handbook of Sociocultural Theory and Second Language Development. London, UK: Routledge, 282–294. https://doi.org/10.4324/9781315624747.

Davin, K. J. and Donato, R. (2013). Student collaboration and teacher-directed classroom dynamic assessment: A complementary pairing. Foreign Language Annals 46 5–22. https://doi.org/10.1111/flan.12012.

Davin, K. J. and Gomez-Pereira, D. (2019). Evaluating instruction through classroom dynamic assessment: A sandwich approach. Language and Sociocultural Theory 6 6–31. https://doi.org/10.1558/lst.38914.

Davin, K. J. , Herazo, J. D. and Sagre, A. (2016). Learning to mediate: Teacher appropriation of dynamic assessment. Language Teaching Research 21 1–20. https://doi.org/10.1177%2F1362168816654309.

Davis, J. M. (2018). Preface. In J. M. Davis , J. M. Norris , M. E. Malone , T. H. McKay and Y. Son (eds.), Useful Assessment and Evaluation in Language Education. Washington, DC: Georgetown University Press, vii–ix. https://doi.org/10.2307/j.ctvvngrq.

Davison, C. and Leung, C. (2009). Current issues in English language teacher-based assessment. TESOL Quarterly 43 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x.

Erben, T. , Ban, R. and Summers, R. (2008). Changing examination structures within a college of education: The application of dynamic assessment in pre-service ESOL endorsement courses in Florida. In J. Lantolf and M. Poehner (eds.), Sociocultural Theory and the Teaching of Second Language. London, UK: Equinox Pub, 87–114. https://doi.org/10.1558/equinox.29294.

Feuerstein, R. , Feuerstein, R. S. , Falik, L. H. and Rand, Y. (2002). The Dynamic Assessment of Cognitive Modifiability: The Learning Propensity Assessment Device: Theory, Instruments and Techniques. Jerusalem, Israel: International Center for the Enhancement of Learning Potential.

Feuerstein, R. , Rand, Y. and Hoffman, M. (1979). The Dynamic Assessment of Retarded Performers: The Learning Potential Assessment Device, Theory, Instruments, and Techniques. Baltimore, MD: University Park Press.

Fulcher, G. and Davidson, F. (eds.) (2012). The Routledge Handbook of Language Testing. London, UK: Routledge. https://doi.org/10.4324/9780203181287.

García, P. N. (2012). Verbalizing in the Second Language Classroom: The Development of the Grammatical Concept of Aspect. Unpublished PhD thesis. University of Massachusetts, Amherst, MA, USA. https://doi.org/10.7275/acvx-px32.

García, P. N. (2014). Verbalizing in the second language classroom: Exploring the role of agency in the internalization of grammatical categories. In P. Deters , X. Gao , E. R. Miller and G. Vitanova (eds.), Theorizing and Analyzing Agency in Second Language Learning: Interdisciplinary Approaches. Bristol, UK: Multilingual Matters, 213–231. https://doi.org/10.21832/9781783092901-014.

García, P. N. (2019). Dynamic assessment: Promoting in-service teachers' conceptual development and pedagogical beliefs in the L2 classroom. Language and Sociocultural Theory 6 32–62. https://doi.org/10.1558/lst.38915.

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention. Journal of Learning Disabilities 42: 111–132.

Grigorenko, E. L. and Sternberg, R. J. (1998). Dynamic testing. Psychological Bulletin 124 75–111. https://doi.org/10.1177/0022219408326207.

Güthke, J. (1982). The learning test concept – an alternative to the traditional static intelligence test. The German Journal of Psychology 6: 306–324.

Haywood, H. and Lidz, C. (2007). Dynamic Assessment in Practice: Clinical and Educational Applications. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511607516.

Hornberger, N. (ed.). (2007). The Encyclopedia of Language and Education, vol. 7: Language Testing and Assessment (E. Shohamy , ed.). Cambridge, UK: Cambridge University Press.

Hill, K. and Sabet, M. (2009). Dynamic speaking assessments. TESOL Quarterly 43 537–545. https://doi.org/10.1002/j.1545-7249.2009.tb00251.x.

Infante, P. and Poehner, M. (2019). Realizing the ZPD in second language education: The complementary contributions of dynamic assessment and mediated development. Language and Sociocultural Theory 6 63–91. https://doi.org/10.1558/lst.38916.
Kozulin, A. and Garb, E. (2002). Dynamic assessment of EFL text comprehension of at-risk students. School Psychology International 23 112–127. https://doi.org/10.1177%2F0143034302023001733.
Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. Journal of Applied Linguistics 1 49–74. http://doi.org/10.1558/japl.v1i1.647.
Lantolf, J. P. and Poehner, M. E. (2011). Dynamic Assessment in the Foreign Language Classroom: A Teacher's Guide, 2nd edn. University Park, PA: Calper.
Lantolf, J. P. and Poehner, M. E. (2013). The unfairness of equal treatment: Objectivity in L2 testing and dynamic assessment. Educational Research and Evaluation 19 141–157. https://doi.org/10.1080/13803611.2013.767616.
Lantolf, J. P. , Poehner, M. E. and Swain, M. (eds.). (2018). Handbook of Sociocultural Theory and Second Language Learning. London, UK: Routledge. https://doi.org/10.4324/9781315624747.
Lantolf, J. and Thorne, S. (2006). Sociocultural Theory and the Genesis of Second Language Development. Oxford, UK: Oxford University Press.
Lee, Y. (2015). Future of diagnostic language assessment. Language Testing 32 295–298. https://doi.org/10.1177/0265532214565385.
Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? Language Assessment Quarterly 4 257–278. https://doi.org/10.1080/15434300701481127.
Levi, T. and Poehner, M. (2018). Employing dynamic assessment to enhance agency among L2 learners. In J. Lantolf , M. Poehner and M. Swain (eds.), The Routledge Handbook of Sociocultural Theory and Second Language Development. London: Routledge, 295–309. https://doi.org/10.4324/9781315624747.
Lidz, C. (1987). Dynamic Assessment. New York, NY: Guilford Press.
Lidz, C. (1991). Practitioner's Guide to Dynamic Assessment. New York, NY: Guilford Press.
Lund, A. (2008). Assessment made visible: Individual and collective practices. Mind, Culture, and Activity 15 32–51. https://doi.org/10.1080/10749030701798623.
McNamara, T. and Roever, C. (2006). Language Testing: The Social Dimension. Oxford, UK: Blackwell Publishing.
Minick, N. (1987). Implications of Vygotsky's theories for dynamic assessment. In C. Lidz (ed.), Dynamic Assessment. New York, NY/: Guilford Press, 116–140.
Negueruela-Azarola, E. , García, P. N. and Buescher, K. (2015). From inter-action to intra-action: The internalization of talk, gesture, and concepts in the second language classroom. In N. Markee (ed.), The Handbook of Classroom Interaction. Malden, MA: Wiley-Blackwell, 233–249. https://doi.org/10.1002/9781118531242.ch14.
Peña, E. and Greene, K. (2018). Dynamic assessment of children learning a second language. In J. Lantolf , M. Poehner and M. Swain (eds.), The Routledge Handbook of Sociocultural Theory and Second Language Development. London, UK: Routledge, 324–340. https://doi.org/10.4324/9781315624747.
Poehner, M. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. Modern Language Journal 91 323–340. https://doi.org/10.1111/j.1540-4781.2007.00583.x.
Poehner, M. (2008a). Both sides of the conversation: The interplay between mediation and learner reciprocity in dynamic assessment. In J. Lantolf and M. Poehner (eds.), Sociocultural Theory and the Teaching of Second Languages. London, UK: Equinox Pub, 33–56. http://doi.org/10.1558/equinox.29292.
Poehner, M. (2008b). Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting l2 Development. Boston, MA: Springer Science.
Poehner, M. (2009). Group dynamic assessment: Mediation for the L2 classroom. TESOL Quarterly 43 471–491. https://doi.org/10.1002/j.1545-7249.2009.tb00245.x.
Poehner, M. (2018). Probing and provoking L2 development: The object of mediation in dynamic assessment and mediated development. In J. Lantolf , M. Poehner and M. Swain (eds.), The Routledge Handbook of Sociocultural Theory and Second Language Development. London, UK: Routledge, 249–265. https://doi.org/10.4324/9781315624747.
Poehner, M. (2019). A Casebook of Dynamic Assessment in Foreign Language Education. University Park, PA: Calper.
Poehner, M. and Infante, P. (2016). Mediated development: A Vygotskian approach to transforming learner L2 abilities. TESOL Quarterly 51 332–357. https://doi.org/10.1002/tesq.308.
Poehner, M. and Lantolf, J. (2005). Dynamic assessment in the language classroom. Language Teaching Research 9 1–33. https://doi.org/10.1191%2F1362168805lr166oa.
Poehner, M. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment. Language Teaching Research 17 323–342. https://doi.org/10.1177%2F1362168813482935.

Poehner, M. and van Compernolle, R. A. (2011). Frames of interaction in dynamic assessment: Developmental diagnoses of second language learning. Assessment in Education: Principles, Policy and Practice 18 183–198. https://doi.org/10.1080/0969594X.2011.567116.
Poehner, M. , Zhang, J. and Lu, X. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. Language Testing 32 337–357. https://doi.org/10.1177%2F0265532214560390.
Rahimi, M. , Kushki, A. and Nassaji, H. (2015). Diagnostic and developmental potentials of dynamic assessment for L2 writing. Language and Sociocultural Theory 2 185–208. https://doi.org/10.1558/lst.v2i2.25956.
Rassaei, E. (2020). Effects of mobile-mediated dynamic and nondynamic glosses on L2 vocabulary learning: A sociocultural Perspective. Modern Language Journal 104: 284–303 https://doi.org/10.1111/modl.12629.
Schneider, E. and Ganschow, L. (2000). Dynamic assessment and instructional strategies for learners who struggle to learn a foreign language. Dyslexia 6 72–82. https://doi.org/10.1002/(SICI)1099-0909(200001/03)6:1%3C72::AID-DYS162%3E3.0.CO;2-B.
Sternberg, R. J. and Grigorenko, E. L. (2002). Dynamic Testing: The Nature and Measurement of Learning Potential. Cambridge, UK: Cambridge University Press.
Turner, C. E. and Purpura, J. E. (2015). Learning-oriented assessment in the classroom. In Handbook of Second Language Assessment. Berlin, Germany and Boston, MA/: DeGruyter Mouton, 255–274.
Tzuriel, D. (1997). The relation between parent-child MLE interactions and children's cognitive modifiability. In A. Kozulin (ed.), The Ontogeny of Cognitive Modifiability. Jerusalem, Israel/: International Center for the Enhancement of Cognitive Modifiability, 157–180.
van Compernolle, R. and Kinginger, C. (2013). Promoting metapragmatic development through assessment in the zone of proximal development. Language Teaching Research 17 282–302. https://doi.org/10.1177%2F1362168813482917.
Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes (M. Cole , V. John-Steiner , S. Scribner and E. Souberman , eds.). Cambridge, MA: Harvard University Press.
Zhang, J. and Lu, X. (2019). Measuring and supporting second language development using computerized dynamic assessment. Language and Sociocultural Theory 6 92–115. https://doi.org/10.1558/lst.31710.

Diagnostic assessment in language classrooms

Alderson, J. C. , Brunfaut, T. and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. Applied Linguistics 36: 236–260. This article reconceptualizes the notion of diagnosis in a broader context across multiple professional fields and offers commonalities applicable to language assessment. Based on interviews with professionals from car mechanics, IT systems support, medicine, psychology, and education, the study discusses five principles of diagnostic language assessment.
Cheng, J. , D'Antilio, Y. Z. , Chen, X. and Bernstein, J. (2014). Automated Assessment of the Speech of Young English Learners. Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. Baltimore, MD: Association for Computational Linguistics, 12–21. This set of proceedings describes the development of a machine learning model to score multilingual children's responses to a speaking test with 14 different item types. The average correlation between human scores and the algorithm (using a 0–4 holistic rating) was .92. This study explains each step of the model development process, including both acoustic and language content modeling, and explains results and limitations by student age, from kindergarten through high school.
Jang, E. E. , Dunlop, M. , Park, G. and van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? Language Testing 32: 359–383. This study examines how holistic diagnostic feedback is interpreted and used by young readers with different cognitive and psychological profiles. The holistic diagnostic feedback consists of information about skill mastery levels, students' self-assessed skill proficiency, and goal orientations. The study results show that students' interpretations and use of feedback for future learning differ significantly by their perceptions about ability and goal orientations. The study highlights that the mechanism of diagnostic feedback use must consider learner characteristics holistically in order to maximise its effects.
Larsen-Freeman, D. (2012). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? Language Teaching 45: 202–214. This article introduces complexity theory as a transdisciplinary theme, which holds a significant implication for applied linguistics. The premise is that the language system can be understood as a system of interdependent systems and parts that interact non-linearly. Language researchers should embrace this complexity by studying modes of system change with a focus on interconnected causes, instead of attempting to isolate individual variables. The author describes 12

principles underlying complexity theory and further delves into dynamism, complexity, and context.
Spolsky, B. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy and A. R. Walton (eds.), Language Assessment for Feedback: Testing and Other Strategies. Dubuque, IA: Kentall, Hunt, 29–41. In this chapter, Spolsky attributes a lack of attention to diagnostic assessment to commitments to proficiency testing devoid of curricular matters. He then discusses various approaches to diagnostic testing from three different perspectives, including pedagogical, linguistic, and processing, as well as conflicts associated with test form and content, the competency-performance divide, and teachers' and students' roles in diagnostic testing.
Alderson, J. C. (2005). Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment. London, UK: Continuum.
Alderson, J. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox , M. Wesche , D. Bayliss , L. Cheng , C. Turner and C. Doe (eds.), Language Testing Reconsidered. Ottawa, Canada: University of Ottawa Press, 21–39. https://doi.org/10.2307/j.ctt1ckpccf.8.
Alderson, J. C. , Brunfaut, T. and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. Applied Linguistics 36 236–260. https://doi.org/10.1093/applin/amt046.
Alderson, J. C. and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. Language Testing 22 301–320. https://doi.org/10.1191/0265532205lt310oa.
Alderson, J. C. and Wall, D. (1993). Does wash back exist? Applied Linguistics 14 115–129. https://doi.org/10.1093/applin/14.2.115.
Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore-Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. International Journal of Listening 35 29–52. https://doi.org/10.1080/10904018.2018.1500915.
Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. Language Testing 30 441–465. https://doi.org/10.1177/0265532212473244.
Beaver, J. (2001). Developmental Reading Assessment K-3 Teacher Resource Guide. Parsippany, NJ: Pearson Learning.
Bejar, I. I. (1984). Educational diagnostic assessment. Journal of Educational Measurement 21 175–189. Retrieved from www.jstor.org/stable/1434541.
Black, P. J. (2009). Formative assessment issues across the curriculum: The theory and the practice. TESOL Quarterly 43 519–523. Retrieved from www.jstor.org/stable/27785033.
Black, P. J. and Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education 5 7–74. https://doi.org/10.1080/0969595980050102.
Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. Language Testing 15 45–85. https://doi.org/10.1177/026553229801500103.
Brunfaut, T. and 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, UK: The British Council. Retrieved from www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf.
Buck, G. and Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. Language Testing 15 119–157. https://doi.org/10.1177/026553229801500201.
Burgin, J. and Hughes, G. D. (2009). Credibly assessing reading and writing abilities for both elementary student and program assessment. Assessing Writing 14: 25–37.
Butler, Y. and Lee, J. (2010). The effects of self-assessment among young learners of English. Language Testing 27 5–31. https://doi.org/10.1177/0265532209346370.
Byrne, S. , Todd, D. , Simpson, B. , Woods, C. and Seidel, R. (2002). Inquiring into learning as system. In G. Ragsdell , D. West and J. Wilby (eds.), Systems Theory and Practice in the Knowledge Age. Boston, MA/: Springer, 115–122.
Byrnes, H. (ed.). (2007). Perspectives. Modern Language Journal 91 641–685. https://doi.org/10.1016/j.asw.2008.12.001.
Chapelle, C. A. , Chung, Y. R. and Xu, J. (eds.). (2008). Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment. Ames, IA: Iowa State University.
Chapelle, C. A. and Douglas, D. (2006). Assessing Language Through Computer Technology. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511733116.
Chen, H. and Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. Language Assessment Quarterly 13 218–230. https://doi.org/10.1080/15434303.2016.1210610.
Cheng, J. , D'Antilio, Y. Z. , Chen, X. and Bernstein, J. (2014). Automated Assessment of the Speech of Young English Learners. Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. Baltimore, MD: Association for Computational Linguistics, 12–21.

Cheng, L. , Watanabe, Y. and Curtis, A. (2004). Washback in Language Testing. Mahwah, NJ: Lawrence Erlbaum Associates.
Cizek, G. J. (ed.). (2001). Setting Performance Standards. Mahwah, NJ: Lawrence Erlbaum Associates.
Cizek, G. J. , Bunch, M. B. and Koons, H. (2004). Setting performance standards: Contemporary methods. Educational Measurement: Issues and Practice 23 31–50. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x.
Clay, M. M. (2000). Running Records for Classroom Teachers. Auckland, NZ: Heinemann Publishers.
Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen , R. Glaser , A. Lesgold and M. Shafto (eds.), Diagnostic Monitoring of Skill and Knowledge Acquisition. Mahwah, NJ/: Lawrence Erlbaum Associates, 75–88.
Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. Annual Review of Applied Linguistics 29 90–100. https://doi.org/10.1017/S0267190509090084.
Davidson, F. and Lynch, B. K. (2002). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. London, UK: Yale University Press.
Dweck, C. S. (1986). Motivational processes affecting learning. American Psychologist 41 1040–1048. https://doi.org/10.1037/0003-066X.41.10.1040.
Edelenbos, P. and Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence.' Language Testing 21 259–283. https://doi.org/10.1191/0265532204lt284oa.
Effatpanah, F. , Baghaei, P. and Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. Language Testing in Asia 9 1–23. https://doi.org/10.1186/s40468-019-0090-y.
Evanini, K. , Heilman, M. , Wang, X. and Blanchard, D. (2015). Automated scoring for the "TOEFL Junior"® comprehensive writing and speaking test: Research report. ETS Research Report Series 1–13. https://doi.org/10.1002/ets2.12052.
Fox, J. D. (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. https://doi.org/10.1016/j.jeap.2008.12.004.
Frederiksen, N. , Glaser, R. , Lesgold, A. and Shafto, M. (eds.). (1990). Diagnostic Monitoring of Skill and Knowledge Acquisition. Mahwah, NJ: Lawrence Erlbaum.
Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment. London, UK: Routledge.
Fulcher, G. (2009). Test architecture. Test retrofit. Language Testing 26 123–144. https://doi.org/10.1177/0265532208097339.
Gibson, A. (2007). CASI Grades 4 to 8 Reading Assessment Teacher's Guide, 2nd edn. Toronto, ON: Thomson Nelson.
Glazer, S. M. and Searfoss, L. W. (1988). Reexamining reading diagnosis. In S. M. Glazer , L. W. Searfoss and L. M. Gentile (eds.), Reexamining Reading Diagnosis: New Trends and Procedures. Newark, NJ/: International Reading Association, 1–11.
Gu, Z. and Jang, E. E. (2009). Investigating the Diagnostic Value of Multiple-Choice Options for Cognitive Diagnostic Assessment. Paper presented at the Canadian Society for the Study of Education (CSSE), Ottawa, Canada.
Harding, L. , Alderson, J. C. and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. Language Testing 32 317–336. https://doi.org/10.1177/0265532214564505.
Henson, R. and Templin, J. (2008). Implementation of Standards Setting for a Geometry End-of-Course Exam. Paper presented at the annual meeting of the American Educational Research Association, New York.
Holton, D. and Clarke, D. (2006). Scaffolding and metacognition. International Journal of Mathematical Education in Science and Technology 37 127–143. https://doi.org/10.1080/00207390500285818.
Hudson, T. and Lynch, B. (1984). A criterion-referenced measurement approach to ESL achievement testing. Language Testing 1 171–210. https://doi.org/10.1177/026553228400100204.
Huff, K. and Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton and M. J. Gierl (eds.), Cognitive Diagnostic Assessment for Education: Theory and Applications. Cambridge, UK: Cambridge University Press, 19–60. https://doi.org/10.1017/cbo9780511611186.
Jang, E. E. (2005). A Validity Narrative: Effects of Reading Skills Diagnosis on Teaching and Learning in the Context of NG TOEFL. Unpublished doctoral dissertation, the University of Illinois, Urbana-Champaign.
Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying fusion model to LanguEdge assessment. Language Testing 26 31–73. https://doi.org/10.1177/0265532208097336.
Jang, E. E. (2010). Implications of Assessment of School Age L2 Students in Ontario. Symposium at Canadian Association of Language Assessment, Montreal, Canada.

Jang, E. E. (2017). Cognitive aspects of language assessment. Language Testing and Assessment 163–177. https://doi.org/10.1007/978-3-319-02326-7_11-1.
Jang, E. E. , Dunlop, M. , Park, G. and van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? Language Testing 32 359–383. https://doi.org/10.1177/0265532215570924.
Jang, E. E. and Ryan, K. (2003). Bridging gaps among curriculum, teaching and learning, and assessment [Review of the book Large-scale assessment: Dimensions, dilemmas, and policy]. Journal of Curriculum Studies 35 499–512. https://doi.org/10.1080/00220270305521.
Jang, E. E. and Sinclair, J. (2019). Advancing Diagnostic Language Assessment Through Natural Language Processing-Based Machine Learning Approaches. Paper presented at the Colloquium, Assessing Young Learners' English Language Proficiency, AAAL, Atlanta, GA.
Jang, E. E. , Vincett, M. , van der Boom, E. H. , Lau, C. and Yang, Y. (2017). Considering young learners' characteristics in developing a diagnostic assessment intervention. In M. K. Wolf and Y. G. Butler (eds.), English Language Proficiency Assessments for Young Learners, vol. 2. New York, NY: Routledge, 193–213. https://doi.org/10.4324/9781315674391-11.
Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. Language Testing 32 227–258. https://doi.org/10.1177/0265532214558457.
Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology 77 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.
Kunnan, A. and Jang, E. E. (2009). Diagnostic feedback in language testing. In M. Long and C. Doughty (eds.), The Handbook of Language Teaching. Oxford, UK: Blackwell Publishing, 610–625. https://doi.org/10.1017/S0261444808005569.
Lantolf, J. P. (2009). Dynamic assessment: The dialectic integration of instruction and assessment. Language Teaching 42 355–368. https://doi.org/10.1017/S0261444808005569.
Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. Journal of Applied Linguistics 1 49–72. https://doi.org/10.1558/japl.v1.i1.49.
Lantolf, J. P. and Poehner, M. E. (2008). Dynamic assessment. In N. H. Hornberger (ed.), Encyclopedia of Language and Education. New York/: Springer, 273–284.
Larsen-Freeman, D. and Cameron, L. (2008). Research methodology on language development from a complex systems perspective. The Modern Language Journal 92 200–213. Retrieved from www.jstor.org/stable/25173023.
Lee, Y. W. , & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. Language Assessment Quarterly 6 172–189. https://doi.org/10.1080/15434300902985108.
Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. Educational Measurement: Issues and Practice 23 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x.
Leighton, J. P. (2009). Mistaken impressions of large-scale cognitive diagnostic testing. In R. P. Phelps (ed.), Correcting Fallacies About Educational and Psychological Testing. Washington, DC: American Psychological Association, 219–246. https://doi.org/10.1037/11861-000.
Leighton, J. P. and Gierl, M. J. (2007). Cognitive Diagnostic Assessment for Education: Theory and Applications. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511611186.
Little, D. (2006). The common European framework of reference for languages: Content, purpose, origin, reception and impact. Language Teaching 39 167–190. https://doi.org/10.1017/S0261444806003557.
McCarthy, A. M. and Christ, T. J. (2010). Test review. In J. M. Beaver and M. A. Carter (eds.), The Developmental Reading Assessment, 2nd edn (DRA2) (2006). Upper Saddle River, NJ: Pearson. Assessment for Effective Intervention 35 182–185. https://doi.org/10.1177/1534508410363127.
McNamara, T. (2000). Language Testing. Oxford, UK: Oxford University Press.
Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols , S. F. Chipman and R. L. Brennan (eds.), Cognitively Diagnostic Assessment. Hillsdale, NJ/: Lawrence Erlbaum Associates Publishers, 43–72.
Mislevy, R. J. , Steinberg, L. S. and Almond, R. G. (2002). Design and analysis in task-based language assessment. Language Testing 19 477–496. https://doi.org/10.1191/0265532202lt241oa.
Mousavi, S. A. (2002). An Encyclopedic Dictionary of Language Testing, 3rd edn. Taiwan: Tung Hua Book Company.
Nichols, P. D. , Meyers, J. L. and Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. Educational Measurement: Issues and Practice 28 14–23. https://doi.org/10.1111/j.1745-3992.2009.00150.x.
Nicols, P. D. , Chipman, S. F. and Brennan, R. L. (eds.). (1995). Cognitively Diagnostic Assessment. Hillsdale, NJ: Lawrence Erlbaum Association, Publishers.

North, B. and Schneider, G. (1998). Scaling descriptors for language proficiency scales. Language Testing 15 217–263. https://doi.org/10.1177/026553229801500204.
Pekrun, R. , Goetz, T. , Titz, W. and Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. Educational Psychologist 37 91–105. https://doi.org/10.1207/S15326985EP3702_4.
Poehner, M. E. and Lantolf, J. P. (2005). Dynamic assessment in the language classroom. Language Teaching Research 9 233–265. https://doi.org/10.1191/1362168805lr166oa.
Ranjbaran, F. and Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. Studies in Educational Evaluation 55 167–179. https://doi.org/10.1016/j.stueduc.2017.10.007.
Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. Journal of Psychoeducational Assessment 34 782–799. https://doi.org/10.1177/0734282915623053.
Ravand, H. and Robitzsch, A. (2015). Cognitive diagnostic modeling using R. Practical Assessment, Research & Evaluation 20 1–12. https://doi.org/10.7275/5g6f-ak15.
Rayner, S. (1998). Educating pupils with emotional and behaviour difficulties: Pedagogy is the key!. Emotional and Behavioural Difficulties 3 39–47. https://doi.org/10.1080/1363275980030206.
Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. Language Testing 18 429–462. https://doi.org/10.1177/026553220101800407.
Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. Journal of Educational Psychology 93 211–222. https://doi.org/10.1037/0022-0663.93.1.211.
Roussos, L. , DiBello, L. , Henson, R. , Jang, E. E. and Templin, J. (2009). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson and J. Roberts (eds.), New Directions in Psychological Measurement with Model-based Approaches. Washington, DC/: American Psychological Association, 35–69.
Rupp, A. , Templin, J. and Henson, R. (2010). Diagnostic Measurement: Theory, Methods, and Applications. New York, NY: Guilford.
Shepard, L. A. (2005). Linking formative assessment to scaffolding. Educational Leadership 63: 66–70.
Sinclair, J. , Jang, E. E. and Rudzicz, F. (2021). Using machine learning to predict children's reading comprehension from linguistic features extracted from speech and writing. Journal of Educational Psychology. (Advance online publication). https://doi.org/10.1037/edu0000658.
Smith, L. B. and Thelen, E. (2003). Development as a dynamic system. Trends in Cognitive Sciences 7 343–348. https://doi.org/10.1016/S1364-6613(03)00156-6.
Spolsky, B. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy and A. R. Walton (eds.), Language Assessment for Feedback: Testing and Other Strategies. Dubuque, IA/: Kentall, Hunt, 29–41.
Stewart, R. (2016). The Psychometric Properties of the Developmental Reading Assessment. Unpublished Master's thesis. Retrieved from https://qspace.library.queensu.ca/bitstream/handle/1974/14823/Stewart_Rachel_A_201608_MED.pdf?sequence=1&isAllowed=y.
Stiggins, G. (1997). Student Centered Classroom Assessment. Upper Saddle River, NJ: Prentice Hall.
Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement 20 345–354. Retrieved from www.jstor.org/stable/1434951.
Toprak, T. E. , Aryadoust, V. and Goh, C. (2019). The 3 log-linear cognitive diagnosis modeling (LCDM) in second language listening assessment. In V. Aryadoust and M. Raquel (eds.), Quantitative Data Analysis for Language Assessment Volume II: Advanced Methods. London, UK: Routledge, 56–78. https://doi.org/10.4324/9781315187808.
Trofimovich, P. , Isaacs, T. , Kennedy, S. , Saito, K. and Crowther, D. (2016). Flawed self-assessment: Investigating self-and other-perception of second language speech. Bilingualism: Language and Cognition 19 122–140. https://doi.org/10.1017/S1366728914000832.
van Geert, P. and van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. Infant Behavior & Development 25 340–374. https://doi.org/10.1016/S0163-6383(02)00140-6.
Vygotsky, L. S. (1978). Socio-cultural theory. Mind in Society 6: 52–58.
Winke, P. M. (2013). The effects of input enhancement on grammar learning and comprehension: A modified replication of Lee (2007) with eye-movement data. Studies in Second Language Acquisition 35 323–352. https://doi.org/10.1017/S0272263112000903.
Woolley, G. (2010). Developing reading comprehension: Combining visual and verbal cognitive processes. Australian Journal of Language and Literacy 33: 108–125.
Yi, Y. S. (2017a). In search of optimal cognitive diagnostic model (s) for ESL grammar test data. Applied Measurement in Education 30 82–101. https://doi.org/10.1080/08957347.2017.1283314.
Yi, Y. S. (2017b). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. Language Testing 34 337–355.

https://doi.org/10.1177/0265532216646141.

Assessing speaking

Fulcher, G. (2003). Testing Second Language Speaking. London: Longman, Pearson Education. This book, although written some time ago, provides a thorough guide to designing and implementing speaking tests, as well as useful and critical summaries of research in L2 speaking assessment.

Fulcher, G. (2015). Assessing second language speaking. Language Teaching 48: 198–216. This article illustrates a timeline of L2 speaking assessment research since the mid-nineteenth century by summarising publications which advanced our understanding of 12 selected themes in the field.

Lim, G. (ed.). (2018). Conceptualizing and operationalizing speaking assessment for a new century [special issue]. Language Assessment Quarterly 15. The articles in this special issue consider important aspects of the speaking construct, such as interactional competence, collocational competence, fluency, and pronunciation and whether and to what extent they have been assessed while reflecting on the potential role of technology in enhancing assessment practices.

Taylor, L. (2011). Examining Speaking: Research and Practice in Assessing Second Language Speaking. Cambridge: UCLES, Cambridge University Press. This edited volume presents a review of relevant literature on the assessment of speaking and provides a systematic framework for validating speaking tests.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance: Assessment in Education: Principles, Policy & Practice 18 279–293. https://doi.org/10.1080/0969594X.2010.526585.

Barkaoui, K. , Brooks, L. , Swain, M. and Lapkin, S. (2012). Test-takers' strategic behaviors in independent and integrated speaking tasks. Applied Linguistics 34 304–324. https://doi.org/10.1093/applin/ams046.

Berry, V. (2007). Personality Differences and Oral Test Performance. Frankfurt: Peter Lang.

Bley-Vroman, R. and Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In A. Mackey and S. Gass (eds.), Research Methodology in Second-Language Acquisition. Hillsdale, NJ/: Lawrence Erlbaum, 245–261.

Boersma, P. and Weenink, D. (2016). Praat: Doing Phonetics by Computer [Computer software]. Retrieved from www.PRAAT.org/.

Bonk, W. J. and Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. Language Testing 20 89–110. https://doi.org/10.1080/15434303.2016.1236797.

Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. Research Notes 11: 20–21.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. Language Testing 26 341–366. https://doi.org/10.1177/0265532209104666.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. Language Testing 20 1–25. https://doi.org/10.1191/0265532203lt242oa.

Brown, A. , Iwashita, N. and McNamara, T. (2005). An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks. TOEFL Monograph No. TOEFL-MS-29. Retrieved from www.ets.org/Media/Research/pdf/RR-05-05.pdf.

Chapelle, C. , Enright, M. and Jamieson, J. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language. London: Routledge.

Chen, L. , Zechner, K. , Yoon, S. , Evanini, K. , Wang, X. , Loukina, A. … Ma, M. (2018). Automated scoring of nonnative speech using the SpeechRater SM v. 5.0 engine. ETS Research Report Series, ETS RR 18–10. https://doi.org/10.1002/ets2.12198.

Davies, L. (2018). Analytic, holistic, and primary trait marking scales. In J. I. Liontas , M. DelliCarpini and TESOL International Association (eds.), The TESOL Encyclopaedia of English Language Teaching, 1–6. https://doi.org/10.1002/9781118784235.eelt0365.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. International Review of Applied Linguistics in Language Teaching 54 113–132. https://doi.org/10.1515/iral-2016-9993.

de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. Language Assessment Quarterly 15 237–254. https://doi.org/10.1080/15434303.2018.1477780.

Ducasse, A. M. and Brown, A. (2011). The role of interactive communication in IELTS Speaking and its relationship to candidates' preparedness for study or training contexts. IELTS Research Reports 12. Retrieved from www.ielts.org/-/media/research-reports/ielts-rr-volume-12-report-3.ashx.

Elder, C. , McNamara, T. , Kim, H. , Pill, J. and Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the nonlanguage specialist can tell us. Language &

Communication 57 14–21. https://doi.org/10.1016/j.langcom.2016.12.005.
Field, J. (2011). Cognitive validity. In L. Taylor (ed.), Examining Speaking: Research and Practice in Assessing Second Language Speaking. Cambridge/: UCLES, Cambridge University Press, 65–111.
Foster, P. and Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. Annual Review of Applied Linguistics 36 98–116. https://doi.org/10.1017/S0267190515000082.
Frost, K. , Elder, C. and Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. Language Testing 29 345–369. https://doi.org/10.1177/0265532211424479.
Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. Language Testing 13 208–238. https://doi.org/10.1177/026553229601300205.
Fulcher, G. (2003). Testing Second Language Speaking. London: Longman, Pearson Education.
Fulcher, G. (2010). Practical Language Testing. London: Hodder Education.
Fulcher, G. (2015a). Assessing second language speaking. Language Teaching 48: 198–216.
Fulcher, G. (2015b). Re-Examining Language Testing: A Philosophical and Social Inquiry. London and New York: Routledge.
Fulcher, G. , Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. Language Testing 28 5–29. https://doi.org/10.1177/0265532209359514.
Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (ed.), Computer-Based Assessment of Foreign Language Speaking Skills. Luxembourg, LU: Publications Office of the European Union, 29–51. Retrieved from www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Volume_European_Commission_2010.pdf.
Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired tests? Applied Linguistics 35 553–574. https://doi.org/10.1093/applin/amt017.
Galaczi, E. D. and Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. Language Assessment Quarterly 15 219–236. https://doi.org/10.1080/15434303.2018.1453816.
Gathercole, S. E. and Baddeley, A. D. (1993). Phonological working memory: A critical building block for reading development and vocabulary acquisition? European Journal of Psychology of Education 8: 259–272.
Green, A. (2012). Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range. Cambridge: UCLES, Cambridge University Press.
Harsch, C. and Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. Assessment in Education: Principles, Policy & Practice 20 281–307. https://doi.org/10.1080/0969594X.2012.742422.
Heritage, J. (1995). Conversation analysis: methodological aspects. In U. M. Quasthoff (ed.), Aspects of Oral Communication. Berlin/: Water de Gruyter, 391–418.
Hsieh, C. and Wang, Y. (2017). Speaking proficiency of young language students: A discourse-analytic study. Language Testing 36 27–50. https://doi.org/10.1177/0265532217734240.
Hutchby, I. and Wooffitt, R. (1998). Conversation Analysis. Cambridge: Cambridge University Press.
Huang, H. T. D. , Hung, S. T. A. and Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. Language Testing 35 27–49. https://doi.org/10.1177/0265532216677106.
Hunter, D. M. , Jones, R. M. and Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. The Canadian Journal of Program Evaluation 11: 61–85.
Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. The Language Learning Journal 44 487–505. https://doi.org/10.1080/09571736.2015.1130079.
Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. Language Assessment Quarterly 15 273–293. https://doi.org/10.1080/15434303.2018.1472264.
Iwashita, N. , McNamara, T. and Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to test design. Language Learning 51 401–436. https://doi.org/10.1111/0023-8333.00160.
Jamieson, J. , Eignor, D. , Grabe, W. and Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle , M. K. Enright and J. M. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language. New York, NY/: Routledge, 55–95.
Johnson, M. (2001). The Art of Non-Conversation: A Re-Examination of the Validity of the Oral Proficiency Interview. New Haven, CT and London: Yale University Press.
Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. Language Testing 34 23–48. https://doi.org/10.1177/0265532215595666.

Khabbazbashi, N. and Galaczi, E. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment Language Testing 37 333–360. https://doi.org/10.1177/0265532219898635.
Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book. London: Longman.
Lam, D. M. K. (2018). What counts as "responding"? Contingency on previous speaker contribution as a feature of interactional competence. Language Testing 35 377–401. https://doi.org/10.1177/0265532218758126.
Latham, H. (1877). On the Action of Examinations Considered as a Means of Selection. Cambridge: Dighton, Bell and Company.
Lazaraton, A. (2002). A Qualitative Approach to the Validation of Oral Language Tests. Cambridge: UCLES, Cambridge University Press.
Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. Language Testing 23 131–166. https://doi.org/10.1191/0265532206lt325oa.
Litman, D. , Strik, H. and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities, Language Assessment Quarterly 15 294–309. https://doi.org/10.1080/15434303.2018.1472265.
May, L. (2011). Interaction in a Paired Speaking Test: The Rater's Perspective. Frankfurt: Peter Lang.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement (3rd edition). London and New York, NY: Palgrave Macmillan.
Morton, H. , Gunson, N. and Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. Advances in Human-Computer Interaction 2012 1–14. https://doi.org/10.1155/2012/389523.
Nakatsuhara, F. (2011). Effects of the number of participants on group oral test performance. Language Testing 28 483–508. https://doi.org/10.1177%2F0265532211398110.
Nakatsuhara, F. (2018). Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests. In G. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam and Philadelphia, PA/: John Benjamins, 205–225.
Nakatsuhara, F. , Inoue, C. , Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. Language Assessment Quarterly 14 1–18. https://doi.org/10.1080/15434303.2016.1263637.
Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics 30 555–578. https://doi.org/10.1093/applin/amp044.
North, B. and Schneider, G. (1998). Scaling descriptors for language proficiency scales. Language Testing 15 217–262. https://doi.org/10.1177/026553229801500204.
Ockey, G. J. (2018). The degree to which it maters if an oral test tasks require listening. In G. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam and Philadelphia, PA/: John Benjamins, 193–204.
Ockey, G. J. , Gu, L. and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. Language Assessment Quarterly 14 346–359. https://doi.org/10.1080/15434303.2017.1400036.
Ockey, G. , Timpe-Laughlin, V. , Davis, L. and Gu, L. (2019). Exploring the potential of a video-mediated interactive speaking assessment. ETS Research Report Series, ETS RR – 19–05. https://doi.org/10.1002/ets2.12240.
O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. Language Testing 36 505–526. https://doi.org/10.1177/0265532219826604.
O'Sullivan, B. (2008). Modelling Performance in Tests of Spoken Language. Frankfurt: Peter Lang.
O'Sullivan, B. and Dunlea, J. (2020). Aptis General technical manual Ver 2.1. Technical Reports, TR/2020/001. Retrieved from www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.1.pdf.
O'Sullivan, B. , Weir, C. J. and Saville, N. (2002). Using observation checklists to validate speaking-test tasks. Language Testing 19 33–56. https://doi.org/10.1191/0265532202lt219oa.
Pallotti, G. (2021). Measuring complexity, accuracy and fluency (CAF). In P. Winke and T. Brunfaut (eds.), The Routledge Handbook of Second Language Acquisition and Language Testing. New York: Routledge, 201–210.
Palmer, H. E. (1921). The Oral Method of Teaching Languages. Cambridge: Heffers.
Plough, I. , Banerjee, J. and Iwashita, N. (2018). Interactional competence: Genie out of the bottle. Language Testing 35 427–445. https://doi:10.1177/0265532218772325.
Pollitt, A. and Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic and N. Saville (eds.), Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium. Cambridge/: UCLES, Cambridge University Press, 74–91.
Purpura, J. E. (2016). Second and foreign language assessment. The Modern Language Journal 100(Supplement) 190–208. https://doi.org/10.1111/modl.12308.

Russell, D. R. (2002). Writing in the Academic Disciplines: A Curricular History, 2nd edn. Carbondale, IL: Southern Illinois University Press.
Sacks, H. , Schegloff, E. and Jefferson, G. (1974). The simplest systematics for the organization of turn-taking for conversation. Language 50 696–735. https://doi.org/10.2307/412243.
Sawaki, Y. , Stricker, L. J. and Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. Language Testing 26 5–30. https://doi.org/10.1177/0265532208097335.
Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. Research on Language and Social Interaction 26 99–128. https://doi.org/10.1207/s15327973rlsi2601_5.
Schegloff, E. A. and Sacks, H. (1973). Opening up closings. Semiotica 8 289–327. http://doi.org/10.1515/semi.1973.8.4.289.
Seedhouse, P. and Nakatsuhara, F. (2018). The Discourse of the IELTS Speaking Test: The Institutional Design of Spoken Interaction for Language Assessment. Cambridge: Cambridge University Press.
Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ/: Educational Testing Service, 89–103.
Sweet, H. (1899). The Practical Study of Languages. London: Dent.
Tavakoli, P. , Nakatsuhara, F. and Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. Modern Language Journal 104 169–191. https://doi.org/10.1111/modl.12620.
Trinity College London . (2021). ISE Rating Scales. www.trinitycollege.com/qualifications/english-language/ISE/ISE-results-and-certificates/ISE-rating-scales
Van Moere, A. (2006). Validity evidence in a university group oral test. Language Testing 23 411–440. https://doi.org/10.1191/0265532206lt336oa.
Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. Language Testing 29 325–344. https://doi.org/10.1177/0265532211424478.
Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. Assessing Writing 9 27–55. https://doi.org/10.1016/j.asw.2004.01.002.
Weir, C. J. , Vidaković, I. and Galaczi, E. D. (2013). Measured Constructs: A History of Cambridge English Language Examinations 1913–2012. Cambridge: UCLES, Cambridge University Press.
Wigglesworth, G. and Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. Language Assessment Quarterly 7 1–24. https://doi.org/10.1080/15434300903031779.
Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. Language Testing 24 251–286. https://doi.org/10.1177/0265532207076365.
Xi, X. , Higgins, D. , Zechner, K. and Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. Language Testing 29 371–394. https://doi.org/10.1177%2F0265532211425673.
Yan, X. , Maeda, Y. , Lv, J. and Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. Language Testing 33 497–528. https://doi.org/10.1177%2F0265532215594643.
Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. Language Testing in Asia 5, Article number 2 1–16. https://doi.org/10.1186/s40468-014-0012-y.

Assessing listening

Buck, G. (2001). Assessing Listening. Cambridge: Cambridge University Press. This book remains the most comprehensive and important work on L2 listening assessment. In the book, Buck provides an overview of L2 listening comprehension, focusing on how listening differs from the other language skills. He also gives his “default listening construct,” which has been hugely influential for listening test developers. He also provides numerous examples of test tasks, suitable texts, illustrative tests of L2 listening, and practical suggestions for creating L2 listening tests.
Geranpayeh, A. and Taylor, L. (eds.). (2013). Examining Listening: Research and Practice in Assessing Second Language Listening. Studies in Language Testing, vol. 35. Cambridge: UCLES, Cambridge University Press. This edited volume is part of the series Studies in Language Testing and examines topics related to the listening components of the Cambridge suite of assessments. (The other books in the series focus on assessing writing, speaking, reading, and young language learners.) Like the other books in the series, this volume follows Weir’s (2005) socio-cognitive framework for language test validation. Of particular interest is John Field’s chapter that examines the issue of cognitive validity in relation to L2 listening.
Ockey, G. and Wagner, E. (2018). Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins. In this hybrid monograph/edited volume, Ockey and Wagner identified four themes related to L2

listening assessment that have been the focus of extensive research: the use of real-world spoken texts, using different types of speech varieties as input, the use of audio-visual texts, and interactive listening as part of the construct of L2 listening ability. The authors provide extensive reviews of the literature for each of these four themes, and then each theme is addressed by two or three empirical studies. They argue that what these four themes have in common is the idea of using more authentic test tasks in order to increase the validity of listening assessments.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice. Oxford, UK: Oxford University Press.

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. Language Testing 13 257–279. https://doi.org/10.1177/026553229601300303.

Batty, A. O. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. Language Testing 32 3–20. https://doi.org/10.1177/0265532214531254.

Batty, A. O. (2018). Investigating the impact of nonverbal communication cues on listening item types. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 161–175. https://doi.org/10.1075/lllt.50.11bat.

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. Language Testing. [OnlineFirst]. https://doi.org/10.1177/0265532220951504.

Bonk, W. J. and Ockey, G. J. (2003). A many-facet Rasch analysis of the L2 group oral discussion task. Language Testing 20 89–110. https://doi.org/10.1191/0265532203lt245oa.

Bradlow, A. R. and Bent, T. (2008). Perceptual adaptation to non-native speech. Cognition 106 707–729. https://doi.org/10.1016/j.cognition.2007.04.005.

Brindley, G. and Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. Language Testing 19 369–394. https://doi.org/10.1191/0265532202lt236oa.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. Language Testing 26 : 341–66. https://doi.org/10.1177/0265532209104666.

Brown, J. D. (2012). Classical test theory. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. Abingdon, UK: Routledge, 323–335. https://doi.org/10.4324/9780203181287.ch22.

Brunfaut, T. (2016). Assessing listening. In D. Tsagari and J. Baneerjee (eds.), Handbook of Second Language Assessment. Boston, MA: De Gruyter, Inc., 97–112. https://doi.org/10.1515/9781614513827-009.

Buck, G. (2001). Assessing Listening. Cambridge: Cambridge University Press.

Burgoon, J. (1994). Nonverbal signals. In M. Knapp and G. Miller (eds.), Handbook of Interpersonal Communication. London, UK/: Routledge, 344–393.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. Language Assessment Quarterly 3 229–242. https://doi.org/10.1207/s15434311laq0303_1.

Carter, R. and McCarthy, M. (2006). Cambridge Grammar of English: A Comprehensive Guide. Spoken and Written English Grammar and Usage. Cambridge, UK: Cambridge University Press.

Chafe, W. (1985). Linguistics differences produced by differences between speaking and writing. In D. Olson , D. Torrance and A. Hildyard (eds.), Literacy Language and Learning. Cambridge, UK/: Cambridge University Press, 105–123.

Choi, I. and So, Y. (2018). A measurement model for listen-speak tasks. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 227–245. https://doi.org/10.1075/lllt.50.15cho.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. System 29 1–14. https://doi.org/10.1016/s0346-251x(00)00057-9.

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. Language Learning and Technology 15: 44–68.

Cublio, J. and Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. Language Assessment Quarterly 10 371–397. https://doi.org/10.1080/15434303.2013.824972.

Douglas, D. (1988). Testing listening comprehension in the context of the ACTFL proficiency guidelines. Studies in Second Language Acquisition 10 345–361. https://doi.org/10.1017/s0272263100007336.

Douglas, D. (1997). Testing Speaking Ability in Academic Contexts: Theoretical Considerations. TOEFL Monograph Series, No. 8. Princeton: Educational Testing Service.

Ericsson, K. and Simon, H. (1993). Protocol Analysis: Verbal Reports as Data, 2nd edn. Cambridge, MA: MIT Press.

Field, J. (2008). Listening in the Language Classroom. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511575945.

Field, J. (2013). Cognitive validity. In A. Geranpayeh and L. Taylor (eds.), Examining Listening: Research and Practice in Assessing Second Language Listening. Studies in Language Testing, vol. 35. Cambridge, UK: UCLES, Cambridge University Press, 77–151.

Fox Tree, J. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. Journal of Memory and Language 34 709–738. https://doi.org/10.1006/jmla.1995.1032.

Freedle, R. and Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. Language Testing 16 2–32. http://dx.doi.org/10.1177/026553229901600102.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired tests? Applied Linguistics 35 553–574. https://doi.org/10.1093/applin/amt017.

Gass, S. M. and Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. Language Learning 34 65–89. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x.

Gruba, P. (1993). A comparison study of audio and video in language testing. JALT Journal 15: 85–88.

Harding, L. (2018). Listening to an unfamiliar accent: Exploring difficulty, strategy use, and evidence of adaptation on listening assessment tasks. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 227–245. https://doi.org/10.1075/lllt.50.07har.

Hilsdon, J. (1995). The group oral exam: Advantages and limitations. In J. Alderson and B. North (eds.), Language Testing in the 1990s: The Communicative Legacy. Hertfordshire, UK/: Prentice Hall International, 189–197.

Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. London, UK: Longman.

Liao, Y. F. , Wagner, E. and Wagner, S. (2018). Test-takers' attitudes and beliefs about the spoken texts used in EFL listening tests. English Teaching & Learning 42 227–236. http://doi.org/10.1007/s42321-018-0013-5.

McCarthy, M. and Carter, R. (1995). Spoken grammar: What is it and how can we teach it? ELT Journal 49 207–218. https://doi.org/10.1093/elt/49.3.207.

Messick, S. (1989). Validity. In R. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: American Council on Education and Macmillan, 13–103.

Messick, S. (1996). Validity and washback in language testing. Language Testing 13 242–256. https://doi.org/10.1177/026553229601300302.

Milroy, J. and Milroy, L. (1999). Authority in Language: Investigating Standard English. London, UK: Routledge.

Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS speaking test. In L. Taylor and C. J. Weir (eds.), IELTS Collected Papers 2: Research in Reading and Listening Assessment. Cambridge, UK/: Cambridge University Press, 519–573.

Ockey, G. J. (2007). Construct implication of including still image or video in computer-based listening tests. Language Testing 24 517–537. https://doi.org/10.1177/0265532207080771.

Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. English for Specific Purposes 35 17–29. https://doi.org/10.1016/j.esp.2013.11.003.

Ockey, G. J. (2018). The degree to which it matters if an oral test task requires listening. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 193–202. https://doi.org/10.1075/lllt.50.13ock.

Ockey, G. J. and Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. Applied Linguistics.

Ockey, G. J. , Koyama, D. , Setoguchi, E. and Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. Language Testing 32 39–62. https://doi.org/10.1177/0265532214538014.

Ockey, G. J. and Wagner, E. (2018a). An overview of interactive listening as part of the construct of interactive and integrated oral test tasks. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 179–192. https://doi.org/10.1075/lllt.50.c12.

Ockey, G. J. (2018b). Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.50.

Oller, J. (1979). Language Tests at School. London, UK: Longman.

Plakans, L. (2012). Writing integrated items. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. Abingdon, UK: Routledge, 249–261. https://doi.org/10.4324/9780203181287.ch17.

Revesz, A. and Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. Studies in Second Language Acquisition 35 31–65. https://doi.org/10.1017/s0272263112000678.

Rost, M. (2015). Teaching and Researching Listening, 3rd edn. London, UK: Taylor and Francis. https://doi.org/10.4324/9781315732862.
Savignon, S. (2018). Communicative competence. In J. Liontas (ed.), The TESOL Encyclopedia of English Language Teaching. Oxford, UK: Wiley-Blackwell. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/9781118784235.eelt0047.
Sueyoshi, A. and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. Language Learning 55 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x.
Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle , H. G. Jun and I. Katz (eds.), Developing and Evaluating Language Learning Materials. Ames, IA/: Iowa State University, 53–68.
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 32 463–483. https://doi.org/10.1177/0265532214562099.
Suvorov, R. (2018). Test-takers' use of visual information in an L2 video-mediated listening test: Evidence from cued retrospective reporting. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 145–160. https://doi.org/10.1075/lllt.50.10suv.
Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (ed.), Spoken and Written Language: Exploring Orality and Literacy. Norwood, NJ/: Ablex, 1–33.
Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), Examining Speaking: Research and Practice in Assessing Second Language Speaking, 30. Cambridge, UK/: Cambridge University Press, 171–233.
Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (eds.), Handbook of Second Language Assessment. Boston, MA: De Gruyter, Inc, 255–272. https://doi.org/10.1515/9781614513827-018.
Wagner, E. (2008). Video listening tests: What are they measuring? Language Assessment Quarterly 5 218–243. https://doi.org/10.1080/15434300802213015.
Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. Language Testing 27 493–513. https://doi.org/10.1177/0265532209355668.
Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. Language Assessment Quarterly 10 178–195. https://doi.org/10.1080/15434303.2013.769552.
Wagner, E. (2014a). Assessing listening. In A. Kunnan (ed.), Companion to Language Assessment, vol. 1. Oxford, UK: Wiley-Blackwell, 47–63. https://doi.org/10.1002/9781118411360.wbcla094.
Wagner, E. (2014b). Using unscripted spoken texts to prepare L2 learners for real world listening. TESOL Journal 5 288–311. https://doi.org/10.1002/tesj.120.
Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banarjee and D. Tsagari (eds.), Contemporary Second Language Assessment. London, UK: Continuum, 438–463. https://doi.org/10.5040/9781474295055.ch-005.
Wagner, E. (2018). Texts for listening instruction and assessment. In J. Liontas (ed.), The TESOL Encyclopedia of English Language Teaching, vol. 3. Oxford, UK: Wiley-Blackwell, 1544–1555. https://doi.org/10.1002/9781118784235.eelt0626.
Wagner, E. and Ockey, G. J. (2018). An overview of the use of audio-visual texts on L2 listening tests. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Amsterdam: John Benjamins, 129–144. https://doi.org/10.1075/lllt.50.c9.
Wagner, E. and Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high stakes tests in China, Japan, and Taiwan. In V. Aryadoust and J. Fox (eds.), Current Trends in Language Testing in the Pacific Rim and the Middle East: Policies, Analyses, and Diagnoses. Newcastle upon Tyne, UK/: Cambridge Scholars Publishing, 103–123.
Wall, D. (2012). Washback. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. Abingdon, UK: Routledge, 79–92. https://doi.org/10.4324/9780203181287.ch5.
Winke, P. and Lim, H. (2014). The Effects of Testwiseness and Test-Taking Anxiety on L2 Listening Test Performance: A Visual (Eye-Tracking) and Attentional Investigation. IELTS Research Reports Online Series 2014/3. Cambridge, UK: British Council, Cambridge English Language Assessment/IDP, IELTS Australia.
Wu, Y. (1998). What do tests of listening comprehension test? – A retrospection study of EFL testtakers performing a multiple-choice task. Language Testing 15 21–44. https://doi.org/10.1191/026553298673885021.
Yanagawa, K. (2016). Examining the authenticity of the center listening test: Speech rate, reduced forms, hesitation and fillers, and processing levels. JACET Journal 60: 97–115.
Zhang, X. (2013). Foreign language listening anxiety and listening performance: Conceptualizations and causal relationships. System 41 164–177. https://doi.org/10.1016/j.system.2013.01.004.

Assessing writing

Weigle, S. C. (2002). Assessing Writing. Cambridge, UK: Cambridge University Press. This book is one of the first volumes entirely devoted to the assessment of writing. It presents a detailed, accessible introduction to various important issues, including constructs, scoring, and task development.

Crusan, D. (2010). Assessment in the Second Language Writing Classroom. Ann Arbor, MI: University of Michigan Press. Deborah Crusan's 2010 book is the only volume devoted to the assessment of writing in classroom contexts. It provides a very accessible introduction for teachers and touches on various assessment types, including portfolio assessment, which is often not discussed in other sources.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing 18: 7–24. This journal article presents a good discussion of the construct assessment by automated essay scoring systems by discussing its usefulness to different types of writing assessment. The article was awarded the best article award by the International Language Testing Association (ILTA) in 2013.

Weigle, S. C. (2013). Assessment of writing. In C. Chapelle (ed.), The Encyclopedia of Applied Linguistics. London, UK: Blackwell Publishing. This encyclopedia chapter provides an excellent short introduction to the assessment of writing, including the nature of second language writing, different types of writing assessments, tasks, and scoring.

Alegria de la Colina, A. and Garcia Mayo, M. P. (2007). Attention to form across collaborative tasks by low-proficiency learners in an EFL setting. In G. Mayo (ed.), Investigating Tasks in Foreign Language Learning. Clevedon, UK/: Multilingual Matters, 91–116.

Bachman, L. and Palmer, A. S. (1996). Language Testing in Practice. Oxford, UK: Oxford University Press.

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. Language Testing 13 257–279. https://doi.org/10.1177/026553229601300303.

Banerjee, J. , Francheschina, F. and Smith, A. M. (2007). Documenting Features of Written Language Production Typical of Different IELTS Band Score Levels. IELTS Research Reports, vol. 7. Retrieved from www.ielts.org/research/research-reports/volume-07-report-5.

Bereiter, C. and Scardamalia, M. (1987). The Psychology of Written Composition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berry, V. (2007). Personality Differences and Oral Test Performance. Frankfurt: Peter Lang.

Bitchener, J. and Ferris, D. (2012). Written Corrective Feedback in Second Language Acquisition. New York, NY: Routledge.

Bitchener, J. and Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. Journal of Second Language Writing 19 207–217. https://doi.org/10.1016/j.jslw.2010.10.002.

Bitchener, J. , Young, S. and Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. Journal of Second Language Writing 14 227–258. https://doi.org/10.1016/j.jslw.2005.08.001.

Bonk, W. and Van Moere, A. (2004). L2 Group Oral Testing: The Influence of Shyness/Outgoingness, Match of Interlocutor's Proficiency Level, and Gender on Indivdiual Scores. Paper presented at the Language Testing Research Colloquium, Temecula.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. Language Testing 26 341–366. https://doi.org/10.1177/0265532209104666.

Brown, J. D. (2016). Introducing Needs Analysis and English for Specific Purposes. Oxford, UK: Routledge.

Byrnes, H. (2014). Theorizing language development at the intersection of 'task' and L2 writing: reconsidering complexity. In H. Byrnes and R. M. Manchon (eds.), Task-Based Language Learning: Insights from and for L2 Writing. London, UK/: John Benjamins, 79–103.

Chapelle, C. , Cotos, E. and Lee, J. Y. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language Testing 32 385–405. https://doi.org/10.1177/0265532214565386.

Cheng, L. , Sun, Y. and Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. Language Teaching 48 436–470. https://doi.org/10.1017/S0261444815000233.

Choi, H. and Iwashita, N. (2016). Interactional behaviours of low-proficiency learners in small group work. In M. Sato and S. Ballinger (eds.), Peer Interaction and Second Language Learning: Pedagogical Potential and Research Agenda. Amsterdam and Philadelphia, PA/: John Benjamins, 113–134.

Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Strasbourg: Council of Europe.

Council of Europe . (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Strasbourg: Council of Europe.

Crusan, D. (2014). Assessing writing. In A. Kunnan (ed.), The Companion to Language Assessment. New York, NY/: John Wiley & Sons, 1–15.

Cumming, A. , Kantor, R. , Baba, K. , Erdosy, U. , Eouanzoui, K. and James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. Assessing Writing 10 1–75. https://doi.org/10.1016/j.asw.2005.02.001.

Danielson, C. and Abrutyn, L. (1997). An Introduction to Using Portfolios in the Classroom. Alexandria: Association for Supervision and Curriculum Development.
Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. Language Testing 26 367–396. https://doi.org/10.1177/0265532209104667.
Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. Language Testing 33 117–135. https://doi.org/10.1177/0265532215582282.
Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing 18 7–24. https://doi.org/10.1016/j.asw.2012.10.002.
Dikli, S. and Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? Assessing Writing 22 1–17. https://doi.org/10.1016/j.asw.2014.03.006.
Eckes, T. , Mueller-Karabil, A. and Zimmermann, S. (2016). Assessing writing. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin/: Walter de Gruyter, 147–164.
Elder, C. and O'Loughlin, K. (2003). Investigating the Relationship Between Intensive EAP Training and Band Score Gains on IELTS. Retrieved from www.ielts.org/research/research-reports/volume-04-report-6.
Elder, C. (2007). ELICOS Language Levels Feasibility Study – Final Report. Melbourne: University of Melbourne, Language Testing Research Centre.
Ellis, R. , Sheen, Y. , Murakami, M. and Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. System 36 353–371. https://doi.org/10.1016/j.system.2008.02.001.
Fernandez Dobao, A. (2012). Collaborative writing tasks in the L2 classroom: Comparing group, pair, and individual work. Journal of Second Language Writing 21 40–58. https://doi.org/10.1016/j.jslw.2011.12.002.
Ferris, D. (2014). Responding to student writing: Teachers' philosophies and practices. Assessing Writing 19 6–23. https://doi.org/10.1016/j.asw.2013.09.004.
Fogal, G. C. (2019). Investigating variability in L2 development: Extending a complexity theory perspective on L2 writing studies and authorial voice. Applied Linguistics 41 575–600. https://doi.org/10.1093/applin/amz005.
Fogal, G. C. and Verspoor, M. H. (eds.). (2020). Complex Dynamic Systems Theory and L2 Writing Development. London, UK: John Benjamins Publishing Company.
Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. Language Testing 13 208–238. https://doi.org/10.1177/026553229601300205.
Fulcher, G. (2010). The reification of the common European framework of reference (CEFR) and effect-driven testing. In A. Psyaltou-Joycey and M. Matthaioudakis (eds.), Advances in Research on Language Acquisition and Teaching. Thessaloniki: GALA.
Fulcher, G. , Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. Language Testing 28 5–29. https://doi.org/10.1177/0265532209359514.
Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? Applied Linguistics 35 553–574. https://doi.org/10.1093/applin/amt017.
Gebril, A. and Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The relationship between discourse features and proficiency. Language Assessment Quarterly 10 9–27. https://doi.org/10.1080/15434303.2011.642040.
Goldstein, L. M. (2006). Feedback and revision in second language writing: Contextual, teacher, and student variables. In K. Hyland and F. Hyland (eds.), Feedback in Second Language Writing: Contexts and Issues. Cambridge, UK/: Cambridge University Press, 185–205.
Grabe, W. and Kaplan, R. B. (1996). Theory and Practice of Writing. New York, NY: Longman.
Guenette, D. and Lyster, R. (2013). Written corrective feedback and its challenges for pre-service ESL teachers. Canadian Modern Language Review 69 129–153. Retrieved from www.muse.jhu.edu/article/506710.
Gurzynski-Weiss, L. K. (2016). Factors influencing Spanish instructors' in-class feedback decisions. The Modern Language Journal 100 255–275. https://doi.org/10.1111/modl.12314.
Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (ed.), Second Language Writing: Research Insights for the Classroom. New York, NY: Cambridge University Press.
Hamp-Lyons, L. (2001). Fourth generation writing assessment. In T. Silva and P. K. Matsuda (eds.), On Second Language Writing. Mahwah, NJ: Lawrence Erlbaum Associates.
Hamp-Lyons, L. and Condon, W. (2000). Assessing the Portfolio: Principles for Practice Theory and Research. Cresskill, NJ: Hampton Press.
Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy and S. Ransdell (eds.), The Science of Writing: Theories, Methods, Individual Differences. Mahwah, NJ, Lawrence Erlbaum Associates.
Hayes, J. R. and Flower, L. S. (1980). Identifying the organization of the writing process. In L. W. Gregg and E. R. Sternberg (eds.), Cognitive Processes in Writing. Hillsdale, NJ/: Erlbaum, 3–30.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. Modern Language Journal 80 309–326. https://doi.org/10.1111/j.1540-4781.1996.tb01613.x.

Hill, K. and McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. Language Testing 29 395–420. https://journals.sagepub.com/doi/pdf/10.1177/0265532211428317.

Hyland, K. (2006). English for Academic Purposes: An Advanced Resource Book. London, UK: Routledge.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. Language Testing 28 509–541. https://doi.org/10.1177/0265532211400860.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy and S. Ransdell (eds.), The Science of Writing: Theories, Methods, Individual Differences, and Applications. Mahwah, NJ: Lawrence Erlbaum.

Knoch, U. (2016, November). Measuring Writing Development: Implications for Research and Pedagogy. Invited keynote presentation at the conference of the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ), University of Auckland, Auckland.

Knoch, U. , Deygers, B. and Khamboonruang, A. (2021). Re-visiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. Language Testing [OnlineFirst]. https://doi.org/10.1177/0265532221994052.

Knoch, U. and Macqueen, S. (2020). Assessing English for Professional Purposes: Language and the Workplace. London, UK: Routledge.

Knoch, U. , Macqueen, S. and O'Hagan, S. (2014). An Investigation of the Effect of Task Type on the Discourse Produced by Students at Various Score Levels in the TOEFL iBT Writing Test. TOEFL iBT Report – 23, ETS Research Report RR-14–43. Princeton, NJ: Princeton University Press.

Knoch, U. , May, L. , Macqueen, S. , Pill, J. and Storch, N. (2016). Transitioning from University to the Workplace: Stakeholder Perceptions of Academic and Professional Writing Demands. IELTS Research Report 2016/1. Retrieved from www.ielts.org/research/research-reports/online-series-2016-1.

Lam, R. (2018). Portfolio Assessment for the Teaching and Learning of Writing. Frankfurt, Germany: Springer.

Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. Journal of Applied Linguistics 1 49–74. Retrieved from https://journals.equinoxpub.com/index.php/JAL/article/view/647.

Leeser, M. J. (2004). Learner proficiency and focus on form during collaborative dialogue. Language Teaching Research 8 55–81. https://doi.org/10.1191/1362168804lr134oa.

Li, J. , Link, S. and Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. Journal of Second Language Writing 27 1–18. https://doi.org/10.1016/j.jslw.2014.10.004.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessments: A longitudinal study of new and experienced raters. Language Testing 28 543–560. https://doi.org/10.1177/0265532211406422.

Long, M. (2005). Methodological issues in learner needs analysis. In M. Long (ed.), Second Language Needs Analysis. Cambridge, UK: Cambridge University Press.

Macqueen, S. (2021). Construct in L2 assessments of speaking. In T. Haug , W. Mann and U. Knoch (eds.), Handbook of Language Assessment Across Modalities. Oxford, UK: Oxford University Press.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. Language Testing 26 397–421. https://doi.org/10.1177/0265532209104668.

McNamara, T. (1996). Measuring Second Language Performance. London and New York, NY: Longman.

Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Palgrave Macmillan, 13–103.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. Journal of Second Language Writing 15 118–141. https://doi.org/10.1016/j.jslw.2006.01.003.

Montgomery, J. L. and Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. Journal of Second Language Writing 16 82–99. https://doi.org/10.1016/j.jslw.2007.04.002.

Nakatsuhara, F. (2004). An Investigation into Conversational Styles in Paired Speaking Tests, MA dissertation. University of Essex, Colchester.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. Language Testing 28 483–508. https://doi.org/10.1177/0265532211398110.

Ockey, G. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. Language Testing 26 161–186. https://doi.org/10.1177/0265532208101005.

O'Loughlin, K. and Arkoudis, S. (2009). Investigating IELTS Exit Score Gains in Higher Education. IELTS Research Reports, vol. 10. Retrieved from www.ielts.org/research/research-reports/volume-10-report-3.

Perry, G. (2019). Comparing Writing and Revisions of Second Language Learners in Computer-Mediated Individual and Collaborative Writing. MA minor thesis. University of Melbourne, Melbourne.
Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on attention and awareness in learning. In R. Schmidt (ed.), Attention and Awareness in Foreign Language Learning. Honolulu, HI/: University of Hawai'i National Foreign Language Resource Center, 1–63.
Schmidt, R. (2001). Attention. In P. Robinson (ed.), Cognition and Second Language Instruction. Cambridge: Cambridge University Press.
Shaw, S. (2004). Creating a virtual community of assessment practice: Towards 'on-line' rater reliability. Cambridge Research Notes 15: 18–20.
Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. TESOL Quarterly 41 255–283. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x.
Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. Assessing Writing 17 81–91. https://doi.org/10.1016/j.asw.2012.02.001.
Storch, N. (2017). Implementing and assessing collaborative writing activities in EAP classes. In J. Bitchener , N. Storch and R. Wette (eds.), Teaching Writing for Academic Purposes to Multilingual Students. New York, NY/: Routledge, 130–142.
Storch, N. and Aldosari, A. (2013). Pairing learners in pair work activity. Language Teaching Research 17 31–48. https://doi.org/10.1177/1362168812457530.
Swain, M. and Lapkin, S. (1995). Problems of output and the cognitive process they generate: A step towards second language learning. Applied Linguistics 16 371–391. https://doi.org/10.1093/applin/16.3.371.
Turner, C. E. and Purpura, J. (2015). Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin, Germany and Boston, MA: DeGruyter Mouton.
Van Beuningen, C. G. , De Jong, N. H. and Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. ITL International Journal of Applied Linguistics 156 279–296. https://doi.org/10.2143/ITL.156.0.2034439.
Verspoor, M. H. and Behrens, H. (2011). Dynamic systems theory and a usage-based approach to second language development. In M. H. Verspoor , K. de Bot and W. Lowie (eds.), A Dynamic Approach to Second Language Development: Methods & Techniques. London, UK/: John Benjamins, 25–38.
Verspoor, M. H. , Lowrie, W. and van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. The Modern Language Journal 92 214–231. https://doi.org/10.1111/j.1540-4781.2008.00715.x.
Wagner, M. (2015). The Centrality of Cognitively Diagnostic Assessment for Advancing Secondary School ESL Students' Writing: A Mixed Methods Study. ProQuest Central; ProQuest Dissertations & Theses Global; Social Science Premium Collection. 1719259603. Retrieved from http://search.proquest.com.ezp.lib.unimelb.edu.au/docview/1719259603?accountid=12372.
Wall, D. (2012). Washback. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. New York, NY/: Routledge, 79–80.
Watts, A. (2006). Fostering Communities of Practice: A Rationale for Developing the Use of New Technologies in Support of Raters. Cambridge, UK: Cambridge University Press.
Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing 15 263–287. https://doi.org/10.1177/026553229801500205.
Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. Assessing Writing 6 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6.
Weigle, S. C. (2002). Assessing Writing. Cambridge, UK: Cambridge University Press.
Wenger, E. (1998). Communities of Practice: Learning, Meaning and Identity. Cambridge, UK: Cambridge University Press.
Wenger, E. and Snyder, W. (2000). Communities of practice: The organizational frontier. Harvard Business Review 78: 139–145.
Wenger, E. , Snyder, W. and McDermott, R. (2002). Cultivating Communities of Practice: A Guide to Managing Knowledge. Boston, MA: Harvard Business School Press.
Williamson, D. , Xi, X. and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice 31 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.
Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. College Composition and Communication 50 483–503. https://doi.org/10.2307/358862.
Yau, M. (1991). The role of language factors in second language writing. In L. Malave and G. Duquette (eds.), Language, Culture and Cognition: A Collection of Studies in First and Second Language Acquisition. Clevedon, UK/: Multilingual Matters, 266–283.
Zhang, Z. and Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. Assessing Writing 36 90–102. https://doi.org/10.1016/j.asw.2018.02.004.

Zupanc, K. and Bosnic, Z. (2015). Advances in the field of automated essay evaluation. Informatica 39: 383–395.

Assessing reading

Alderson, J. C. (2000). Assessing Reading. Cambridge: Cambridge University Press. This remains one of the most comprehensive handbooks available on the testing of L2 reading, offering both theoretical and practical insights. Starting with an accessible review of the construct of reading and key factors that impact it, Alderson then discusses implications for the testing of reading. The handbook describes different approaches to operationalising reading constructs and designing reading tests, thereby giving due attention to the reader, the target language use domain, text and task characteristics, and test purposes. Practitioners might particularly value the overview of potential task types for assessing L2 reading in Chapter 7, including the many example tasks and the outline of their pros and cons.

Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (2008). Building a validity argument for the Test of English as a Foreign Language. New York, NY: Routledge. This resource provides an example of a framework and guidelines for validating reading assessments and also illustrates, by means of a widely used reading test, how this can be done in practice. Namely, Chapelle *et al*. demonstrate the application of the argument-based validation approach to the TOEFL reading test (as well as other skills sections of this test).

Grabe, W. and Stoller, F. L. (2020). Teaching and Researching Reading, 3rd edn. Oxon and New York, NY: Routledge. Now in its third edition, Grabe and Stoller's volume is a key resource for anyone working in the area of second/foreign language reading, whether as a researcher or a practitioner. For language testers, Parts I and II are must-reads; these provide in-depth insights into L1 and L2 reading from both theoretical and empirical perspectives and thus essentially describe the construct we are aiming to assess. The review is well-referenced and therefore offers many leads for further reading, as does the final "resources" chapter. The concept boxes presented throughout the chapters reinforce important points and act as summaries to which one can easily return.

Khalifa, H. and Weir, C. J. (2009). Examining Reading. Cambridge: Cambridge University Press. This book provides another worked example of how to validate reading assessments using an established framework. More specifically, Khalifa and Weir describe the use of the socio-cognitive approach to evaluate the validity of the Cambridge ESOL Main Suite's reading tests (now called Cambridge English Qualifications).

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. Language Testing 24 7–36. https://doi.org/10.1177/0265532207071510.

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson and A. H. Urquhart (eds.), Reading in a Foreign Language. London: Longman.

Alderson, J. C. (2000). Assessing Reading. Cambridge: Cambridge University Press.

Alderson, J. C. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: Cambridge University Press.

Anderson, N. J. , Bachman, L. , Perkins, K. and Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. Language Testing 8 41–66. https://doi.org/10.1177/026553229100800104.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. Language Testing 28 51–75. https://doi.org/10.1177/0265532210376379.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. Language Testing 30 441–465. https://doi.org/10.1177/0265532212473244.

Bax, S. and Chan, S. (2016). Researching the Cognitive Validity of GEPT High-Intermediate and Advanced Reading: An Eye Tracking and Stimulated Recall Study. Taiwan: The Language Training and Testing Center.

Bowles, M. A. (2010a). The Think-Aloud Controversy in Second Language Research. New York and Oxon: Routledge.

Bowles, M. A. (2010b). Concurrent verbal reports in second language acquisition research. Annual Review of Applied Linguistics 30 111–127. https://doi.org/10.1017/S0267190510000036.

Brunfaut, T. , Kormos, J. , Michel, M. and Ratajczak, M. (2021). Testing young foreign language learners' reading comprehension: Exploring the effects of working memory, grade level and reading task. Language Testing 38 356–377. https://doi.org/10.1177/0265532221991480.

Brunfaut, T. and McCray, G. (2015a). Looking into Test-Takers' Cognitive Processing Whilst Completing Reading Tasks: A Mixed-Methods Eye-Tracking and Stimulated Recall Study. ARAGs Research Reports Online, AR/2015/01. London: British Council. Retrieved from www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final.pdf.

Brunfaut, T. and McCray, G. (2015b). Modelling L2 Reading Proficiency Using Eye-Tracking Measures. Paper presented at the CRELLA/LU Eye tracking Research Day, Luton, UK.

Burton, S. J. , Sudweeks, R. R. , Merrill, P. F. and Wood, B. (1991). How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty. Brigham Young University Testing Services and The Department of Instructional Science. Retrieved from https://testing.byu.edu/handbooks/betteritems.pdf.

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. Scientific Studies of Reading 1 3–43. https://doi.org/10.1207/s1532799xssr0101_2.

Chan, S. H. C. (2013). Establishing the Validity of Reading-into-Writing Tasks for the UK Academic Context. Unpublished doctoral dissertation. University of Bedfordshire, Luton, UK.

Chan, S. H. C. (2018). Defining Integrated Reading-into-Writing Constructs: Evidence at the B2 C1 Interface. English Profile Series Studies 08. Cambridge: Cambridge University Press.

Chan, S. H. C. , Wu, R. Y. F. and Weir, C. J. (2014). Examining the Context and Cognitive Validity of the GEPT Advanced Writing Task 1: A Comparison with Real-Life Academic Writing Tasks. LTTC-GEPT Research Reports, RG-03. Taiwan: Language Training and Testing Center.

Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (2008). Building a Validity Argument for the Test of English as a Foreign Language. New York, NY: Routledge.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. Review of Educational Research 49 222–251. https://doi.org/10.3102/00346543049002222.

Enright, M. K. , Grabe, W. , Koda, K. , Mosenthal, P. B. , Mulcahy-Ernt, P. and Schedl, M. A. (2000). TOEFL 2000 Reading Framework: A Working Paper. TOEFL Monograph Series, TOEFL-MS-17. Princeton, NJ: Educational Testing Service.

Francis, D. J. , Kulesz, P. A. and Benoit, J. S. (2018). Extending the simple view of reading to account for variation within readers and across texts: The complete view of reading (CVRi). Remedial and Special Education 39 274–288. https://doi.org/10.1177/0741932518772904.

Freedle, R. and Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. Language Testing 10 133–170. https://doi.org/10.1177/026553229301000203.

Gass, S. M. and Mackey, A. (2017). Stimulated Recall Methodology in Applied Linguistics and L2 Research. New York and Oxon: Routledge.

Gebril, A. and Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. Assessing Writing 21 56–73. https://doi.org/10.1016/j.asw.2014.03.002.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. Journal of the Reading Specialists 24 44–69. https://doi.org/10.1080/19388076709556976.

Gough, P. B. (1972). One second of reading. Visible Language 6: 291–320.

Grabe, W. (2009). Reading in a Second Language: Moving from Theory to Practice. New York, NY: Cambridge University Press.

Grabe, W. and Jiang, X. (2014). Assessing reading. In A. J. Kunnan (ed.), The Companion to Language Assessment. Hoboken, NJ/: Wiley-Blackwell, 185–200.

Grabe, W. and Stoller, F. L. (2020). Teaching and Researching Reading, 3rd edn. Oxon and New York, NY: Routledge.

Green, A. (1998). Verbal Protocol Analysis in Language Testing Research: A Handbook. Cambridge: Cambridge University Press.

Green, R. (2013). Statistical Analyses for Language Testers. London: Palgrave Macmillan.

Green, R. (2017). Designing Listening Tests: A Practical Approach. London: Palgrave Macmillan.

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. Reading Research Quarterly 23 178–199. https://doi.org/10.2307/747801.

In'nami, Y. and Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. Language Testing 26 219–244. https://doi.org/10.1177/0265532208101006.

Jamieson, J. , Jones, S. , Kirsch, I. , Mosenthal, P. and Taylor, C. (1999). TOEFL 2000 Framework: A Working Paper. TOEFL Monograph Series, TOEFL-MS-16. Princeton, NJ: Educational Testing Service.

Jeon, E. H. and Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. Language Learning 64 160–212. https://doi.org/10.1111/lang.12034.

Kamata, A. and Vaughn, B. K. (2004). An introduction to differential item functioning analysis. Learning Disabilities: A Contemporary Journal 2: 49–69.

Khalifa, H. and Weir, C. J. (2009). Examining Reading. Cambridge: Cambridge University Press.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. Psychological Review 95: 163–182.

Kremmel, B. , Brunfaut, T. and Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. Applied Linguistics 38 848–870. https://doi.org/10.1093/applin/amv070.

Koo, J. , Becker, B. J. and Kim, Y-S . (2013). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. Language Testing 31 89–109. https://doi.org/10.1177/0265532213496097.

Löwenadler, J. (2019). Patterns of variation in the interplay of language ability and general reading comprehension ability in L2 reading. Language Testing 36 369–390. https://doi.org/10.1177/0265532219826379.
McCray, G. (2014). Statistical Modelling of Cognitive Processing in Reading Comprehension in the Context of Language Testing. Unpublished doctoral dissertation. Lancaster University, Lancaster, UK.
McCray, G. and Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. Language Testing 35 51–73. https://doi.org/10.1177/0265532216677105.
Mundy, J. (1968). Read and Think. Harlow: Longman.
Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. Language Learning 23 105–118. https://doi.org/10.1111/j.1467-1770.1973.tb00100.x.
O'Sullivan, B. and Dunlea, J. (2015). Aptis General Technical Manual, Version 1.0. Technical Report, TR/2015/005. London: British Council.
O'Sullivan, B. and Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (ed.), Language Testing: Theories and Practices. Basingstoke/: Palgrave Macmillan, 13–32.
Pae, T. I. (2004). DIF for examinees with different academic backgrounds. Language Testing 21 53–73. https://doi.org/10.1191/0265532204lt274oa.
Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. Language Testing 29 533–554. https://doi.org/10.1177/0265532211434027.
Pae, T. I. (2019). A simultaneous analysis of relations between L1 and L2 skills in reading and writing. Reading Research Quarterly 54 109–124. https://doi.org/10.1002/rrq.216.
Perfetti, C. A. (1997). Sentences, individual differences, and multiple texts: Three issues in text comprehension. Discourse Processes 23 337–355. https://doi.org/10.1080/01638539709544996.
Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. Language Testing 20 26–56. https://doi.org/10.1191/0265532203lt243oa.
Plakans, L. and Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. Assessing Writing 17 18–34. https://doi.org/10.1016/j.asw.2011.09.002.
Plakans, L. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. Journal of Second Language Writing 22 217–230. https://doi.org/10.1016/j.jslw.2013.02.003.
Raatz, U. and Klein-Braley, C. (1981). The C-test – a modification of the cloze procedure. In T. Culhane , C. Klein-Braley and D. K. Stevenson (eds.), Practice and Problems in Language Testing. University of Essex Department of Language and Linguistics Occasional Papers, 26. Colchester: University of Essex.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124 372–422. https://doi.org/10.1037/0033-2909.124.3.372.
Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. Language Testing in Asia 3. https://doi.org/10.1186/2229-0443-3-9.
Rupp, A. A. , Ferne, T. and Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. Language Testing 23 441–474. https://doi.org/10.1191/0265532206lt337oa.
Sarig, G. (1989). Testing meaning construction: Can we do it fairly? Language Testing 6 77–94. https://doi.org/10.1177/026553228900600107.
Smith, F. (1971). Understanding Reading. New York, NY: Holt, Rinehart & Winston.
Stanovich, K. E. (1980). Toward an interactive compensatory model of individual differences in the development of reading fluency. Reading Research Quarterly 16 32–71. https://doi.org/10.2307/747348.
Tunmer, W. E. and Chapman, J. W. (2012). The simple view of reading redux vocabulary knowledge and the independent components hypothesis. Journal of Learning Disabilities 45 453–466. https://doi.org/10.1177/0022219411432685.
Wang, J. , Engelhard, G. , Raczynski, K. , Song, T. and Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. Assessing Writing 33 36–47. https://doi.org/10.1016/j.asw.2017.03.003.
Weigle, S. C. and Montee, M. (2012). Raters' perception of textual borrowing in integrated writing tasks. In E. E. Van Steendam , M. Tillema , G. Rijlaarsdam and H. Van Den Bergh (eds.), Measuring Writing: Recent Insights into Theory, Methodology and Practises. London/: Brill, 117–145.
Weir, C. J. (2005). Language Testing and Validation. Basingstoke: Palgrave Macmillan.
Winke, P. M. (2014). Eye-tracking technology for reading. In A. J. Kunnan (ed.), The Companion to Language Assessment. Hoboken, NJ: Wiley-Blackwell, 1029–1046.

Test specifications

Davidson, F. and Lynch, B. (2002). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. New Haven, CT and London: Yale University Press. This book is about language test development using test specifications. Viewed as the basic tool of testcraft, test specifications are introduced and discussed from various perspectives, including their role in specification-driven, inclusive testing.

Fulcher, G. (2015). Re-Examining Language Testing – A Philosophical and Social Inquiry. London and New York: Routledge. This monograph provides a unique analysis of the ideas and social forces that shape the practice of language testing and assessment. The notion of Pragmatic realism proposed in the monograph is enlightening and thought provoking, lending strong support to effect-driven testing.

Fulcher, G. and Davidson, F. (2009). Test architecture, test retrofit. Language Testing 26: 123–144. This article emphasizes the central role of test purposes in language test design and revision and presents a systematic approach to evaluating test revisions or retrofit. The use of the architecture metaphor makes the argument accessible to a broad audience including language testing practitioners.

Alderson, C. J. , Brunfaut, T. and Harding, L. (2017). Bridging assessment and learning: A view from second and foreign language assessment, Assessment in Education: Principles, Policy & Practice 24 379–387. https://doi.org/10.1080/0969594X.2017.1331201.

Alderson, C. J. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: Cambridge University Press.

Alderson, C. J. , Haapakangas, E. L. , Huhta, A. , Nieminen, L. and Ullakonoja, R. (2015). The Diagnosis of Reading in a Second or Foreign Language. New York: Routledge.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L. F. and Damböck, B. (2018). Language Assessment for Classroom Teachers. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press.

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

Carroll, J. B. (1961). Fundamental Considerations in Testing for English Language Proficiency of Foreign Students. Washington, DC: Testing Center for Applied Linguistics.

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

Cheng, L. and Curtis, A. (eds.). (2010). English Language Assessment and the Chinese Learner. London: Routledge.

Cizek, G. J. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (ed.), Defending Standardized Testing. Mahwah, NJ/: Lawrence Erlbaum Associates, Inc., 23–54.

The College Board . (2015). Test Specifications for the Redesigned SAT®. Retrieved August 15, 2019, from www.collegeboard.org.

Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching and Assessment. Strasbourg: Council of Europe.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London and New York/: Routledge, 197–207.

Davidson, F. and Lynch, B. (2002). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. New Haven, CT and London: Yale University Press. Retrieved from www.jstor.org/stable/j.ctt1npx6r.

Davies, A. , Brown, A. , Elder, C. , Hill, K. , Lumley, T. and McNamara, T. (1999). Dictionary of Language Testing. Cambridge: Cambridge University Press.

DELTA . (2018). Guidelines for Users. Retrieved October 15, from http://gslpa.polyu.edu.hk/eng/delta_web/.

Fulcher, G. (2010a). Practical Language Testing. London: Hodder Education.

Fulcher, G. (2010b). The reification of the common European framework of reference (CEFR) and effect-driven testing. In A. Psyaltou-Joycey and M. Matthaioudakis (eds.), Advances in Research on Language Acquisition and Teaching. Thessaloniki, Greece/: GALA, 15–26.

Fulcher, G. (2015). Re-Examining Language Testing – A Philosophical and Social Inquiry. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment – An Advanced Resource Book. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2009). Test architecture, test retrofit. Language Testing 26: 123–144.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist 18 519–521. https://doi.org/10.1037/h0049294.

Green, A. (2015). The ABCs of assessment. In D. Tsagari , K. Vogt , V. Froehlich , I. Csépes , A. Fekete , A. Green … S. Kordia (eds.), Handbook of Assessment for Language Teachers, 1–15. Retrieved November 11, 2019, from http://taleproject.eu.

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperez and D. Hymes (eds.), Directions in Sociolinguistics. New York/: Holt, Rinehart and Winston, 35–71.

Jin, Y. and Zhang, L. (2016). The impact of test mode on the use of communication strategies in paired discussion. In G. Yu and Y. Jin (eds.), Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums. Basingstoke/: Palgrave Macmillan, 61–84.

Kim, J. and Davidson, F. (2014). Effect-driven test specifications. In A. J. Kunnan (ed.), The Companion to Language Assessment, II:7:47. Chichester, UK/: Wiley Blackwell, 788–795.

Lee, Y. W. (2015). Future of diagnostic language assessment. Language Testing 32 295–298. https://doi.org/10.1177/0265532214565385.

Leung, C. (2005). Classroom teacher assessment of second language development: Construct as practice. In E. Hinkel (ed.), Handbook of Research in Second Language Teaching and Learning. Mahwah, NJ/: Lawrence, 869–888.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. Language Testing 17 43–64. https://doi.org/10.1177/026553220001700102.

Messick, S. (1996). Validity and washback in language testing. Language Testing 13 241–256. https://doi.org/10.1177/026553229601300302.

Mislevy, R. J. (2003a). On the Structure of Educational Assessments. CSE Technical Report 597. Los Angeles: Centre for the Study of Evaluation, CRESST.

Mislevy, R. J. (2003b). Argument Substance and Argument Structure in Educational Assessment. CSE Technical Report 605. Los Angeles: Centre for the Study of Evaluation, CRESST.

Mislevy, R. J. , Almond, R. G. and Lukas, J. F. (2003). A Brief Introduction to Evidence-Centred Design. Research Report RR-03-16. Princeton, NJ: Educational Testing Service.

National College English Testing Committee . (2016). College English Test Syllabus, revised version. Shanghai: Shanghai Jiao Tong University Press.

Oller, J. (1979). Language Tests at School. London: Longman.

O'Sullivan, B. and Green, A. (2011). Test taker characteristics. In L. Taylor (ed.), Examining Speaking: Research and Practice in Assessing Second Language Speaking. Studies in Language Testing 30. Cambridge: UCLES, Cambridge University Press, 36–64.

O'Sullivan, B. and Weir, C. J. (2011). Language testing and validation. In B. O'Sullivan (ed.), Language Testing: Theory and Practices. Basingstoke/: Palgrave Macmillan, 13–32.

Ruch, G. M. (1929). The Objective or New-Type Examination: An Introduction to Educational Measurement. Chicago: Scott, Foresman.

Shohamy, E. and McNamara, T. (2009). Language tests for citizenship, immigration, and asylum. Language Assessment Quarterly 6 1–5. https://doi.org/10.1080/15434300802606440.

Spaan, M. (2006). Test and item specifications development. Language Assessment Quarterly 3: 71–79.

Spolsky, B. (1995). Measured Words. Oxford: Oxford University Press.

Tan, M. and Turner, C. E. (2015). The impact of communication and collaboration between test developers and teachers on a high-stakes ESL exam: Aligning external assessment and classroom practices. Language Assessment Quarterly 12 29–49. https://doi.org/10.1080/15434303.2014.1003301.

Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), Examining Speaking: Research and Practice in Assessing Second Language Speaking. Studies in Language Testing, vol. 30. Cambridge: UCLES, Cambridge University Press, 171–233.

Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin/: DeGruyter, 255–272.

Vogt, K. and Froehlich, V. (2015). Alternatives in assessment. In D. Tsagari , K. Vogt , V. Froehlich , I. Csépes , A. Fekete , A. Green … S. Kordia (eds.), Handbook of Assessment for Language Teachers, 148–178. Retrieved November 11, 2019, from http://taleproject.eu.

Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke: Palgrave Macmillan.

Wiggins, G. P. (1998). Educative Assessment: Designing Assessments to Inform and Improve Student Performance. San Francisco: Jossey-Bass Publishers.

Xerri, D. and Briffa, P. V. (eds.). (2018). Teacher Involvement in High-Stakes Language Testing. Switzerland: Springer International Publishing AG.

Yu, G. and Jin, Y. (eds.). (2016). Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums. Basingstoke: Palgrave Macmillan.

Evidence-centered design in language testing

Arieli-Attali, M. , Ward, S. , Thomas, J. , Deonovic, B. and von Davier, A. A. (2019). The expanded evidence-centered-design (e-ECD) for learning and assessment systems: A framework to incorporating learning goals and processes within assessment design. Frontiers in Psychology 10: 853. This article extends the ECD framework to learning. A model for learning is used to provide coordinated extensions to the ECD student, task, and evidence models. It is applicable when learning occurs both with and without explicit instruction.

Chapelle, C. , Enright, M. K. and Jamieson, J. M. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language TM. London: Routledge. This volume presents an in-depth discussion of the application of some of the ideas of evidence-centered design and assessment use argumentation to a TOEFL redesign.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. Journal of Applied Testing Technology 14: 1–38. The assessment engineering framework is compatible with ECD but provides additional structural support for task design, scoring, and psychometric modeling schemas.

Mislevy, R. J. (2018). Sociocognitive Foundations of Educational Measurement. New York & London: Routledge. This monograph is an in-depth interdisciplinary treatise of foundational issues in assessment and measurement as re-examined from a socio-cognitive perspective.

Mislevy, R. J. , Almond, R. G. and Lukas, J. (2004). A Brief Introduction to Evidence-Centered Design. CSE Technical Report 632. The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. www.cse.ucla.edu/products/reports/r632.pdf. This is a reader-friendly introduction to evidence-centered design. It is the best place for a new reader to start.

Mislevy, R. J. , Steinberg, L. S. and Almond, R. A. (2002). Design and analysis in task-based language assessment. Language Testing 19: 477–496. This is a more technical discussion of the application of evidence-centered design to task-based language tests. It addresses issues of construct definition, multivariate student models, and complex tasks.

Mislevy, R. J. and Yin, C. (2009). If language is a complex system, what is language assessment? Language Learning 59 (Supplement 1): 249–267. This article discusses how to build and interpret assessment arguments from an interactionist/complex socio-cogntive systems perspective on language.

Riconscente, M. M. , Mislevy, R. J. and Corrigan, S. (2015). Evidence-centered design. In S. Lane, T. M. Haladyna and M. Raymond (eds.), Handbook of Test Development. London: Routledge, 40–63. This discussion of ECD is aimed at test developers. It uses representations and terminology from the Principled Assessment Design for Inquiry (PADI) project.

Shute, V. J. , Masduki, I. and Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. Technology, Instruction, Cognition & Learning 8: 137–161. In a series of articles and projects, Shute and her colleagues have used ECD in designing and implementing unobtrusive assessment in the context of games and simulations. This article explains and illustrates the approach.

Wilson, M. (2005). Constructing Measures. Mahwah, NJ: Lawrence Erlbaum. This monograph is consistent with the ECD framework but goes more deeply into the definition, task modeling, and Rasch psychometric modeling of constructs. It is well suited for assessment built on learning progressions.

Almond, R. , Steinberg, L. and Mislevy, R. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. The Journal of Technology, Learning and Assessment 1. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1671.

Arieli-Attali, M. , Ward, S. , Thomas, J. , Deonovic, B. and von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. Frontiers in Psychology 10. https://doi.org/10.3389/fpsyg.2019.00853.

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. The Modern Language Journal 86 525–545. https://doi.org/10.1111/1540-4781.00159.

Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quarterly 2 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox , M. Wesche and D. Bayliss (eds.), Language Testing Reconsidered. Ottowa, ON: University of Ottawa Press.

Bachman, L. F. and Cohen, A. D. (eds.). (1998). Interfaces Between Second Language Acquisition and Language Testing Research. Cambridge: Cambridge University Press.

Bachman, L. F. and Palmer, A. S. (2010). Language Assessment Practice: Developing Language Assessments and Justifying Their Use in the Real World. Oxford: Oxford University Press.

Behrens, J. T. , Mislevy, R. J. , DiCerbo, K. E. and Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In C. M. Michael , J. Clarke-Midura and D. H. Robinson (eds.), Technology-Based Assessments for 21st Century Skills. Charlotte, NC: Information Age Publishing.

Bloom, B. S. (ed.). (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I, Cognitive Domain. London: Longman.
Bormuth, J. R. (1970). On the Theory of Achievement Test Items. Chicago: University of Chicago Press.
Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. Language Testing 20 369–383. https://doi.org/10.1191/0265532203lt264oa.
Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), Interfaces Between Second Language Acquisition and Language Testing Research. Cambridge: Cambridge University Press.
Chapelle, C. , Enright, M. K. and Jamieson, J. M. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language TM. London: Routledge.
College Board . (n.d.). AP Course and Exam Redesign. Retrieved from https://aphighered.collegeboard.org/courses-exams/course-exam-redesign.
Conrad, S. , Clarke-Midura, J. and Klopfer, E. (2014). A framework for structuring learning assessment in an educational massively multiplayer online educational game – experiment centered design. International Journal of Game-Based Learning 4 37–59. https://doi.org/10.4018/IJGBL.2014010103.
Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching and Assessment. Cambridge University Press. Retrieved from www.coe.int/t/dg4/linguistic/source/Framework_EN.pdf.
Davidson, F. and Lynch, B. K. (2001). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. New Haven, CT: Yale University Press.
Douglas, D. (2000). Assessing Languages for Specific Purposes. Cambridge: Cambridge University Press.
Embretson, S. E. (ed.). (1985). Test Design: Developments in Psychology and Psychometrics. New York: Academic Press.
Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods 3 380–396. https://doi.org/10.1037/1082-989X.3.3.380.
Feng, M. , Hansen, E. G. and Zapata-Rivera, D. (2009, April 13–17). Using Evidence Centered Design for Learning (ECDL) to Examine the Assistments System. Paper presentation. American Educational Research Association Annual Meeting, San Diego, CA.
Fulcher, G. and Davidson, F. (2009). Test architecture: Test retrofit. Language Testing 26 123–144. https://doi.org/10.1177/0265532208097339.
Gierl, M. J. , Zhou, J. and Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. Journal of Technology, Learning, and Assessment 7. Retrieved from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1137&context=jtla.
Grover, S. , Bienkowski, M. , Basu, S. , Eagle, M. , Diana, N. and Stamper, J. (2017). A Framework for Hypothesis-Driven Approaches to Support Data-Driven Learning Analytics in Measuring Computational Thinking in Block-Based Programming. Seventh International Learning Analytics & Knowledge Conference (ACM), New York, NY.
Haladyna, T. M. and Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. Evaluation and the Health Professions 12 97–104. https://doi.org/10.1177/016327878901200106.
Higgins, D. , Xi, X. , Zechner, K. and Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. Computer Speech & Language 25 282–306. https://doi.org/10.1016/j.csl.2010.06.001.
Hines, S. (2010). Evidence-Centered Design: The TOEIC® Speaking and Writing Tests. Educational Testing Service. www.ets.org/research/policy_research_reports/publications/report/2010/itjx.
Hively, W. , Patterson, H. L. and Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement 5 275–290. Retrieved from www.jstor.org/stable/1433778.
Irvine, S. H. and Kyllonen, P. C. (eds.). (2002). Item Generation for Test Development. Hillsdale, NJ: Lawrence Erlbaum Associates.
Kane, M. (2006). Validation. In R. J. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: Praeger.
Kane, M. (2013). Validating the interpretation and use of test scores. Journal of Educational Measurement 50 1–73. https://doi.org/10.1111/jedm.12000.
Ke, F. , Shute, V. , Clark, K. M. and Erlebacher, G. (2019). Interdisciplinary Design of Game-Based Learning Platforms. Heidelberg, Germany: Springer Nature.
Larsen-Freeman, D. and Cameron, L. (2008). Complex Systems and Applied Linguistics. Oxford: Oxford University Press.
Lee, Y. W. and Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. Language Assessment Quarterly 6 172–189. https://doi.org/10.1080/15434300902985108.
Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. Educational Measurement: Issues and Practice 27 32–42. https://doi.org/10.1111/j.1745-3992.2008.00126.x.

Long, M. H. (2015). Second Language Acquisition and Task-Based Language Teaching. Oxford: Wiley-Blackwell.
Luecht, R. M. (2003). Multistage complexity in language proficiency assessment: A framework for aligning theoretical perspectives, test development, and psychometrics. Foreign Language Annals 36 527–535. https://doi.org/10.1111/j.1944-9720.2003.tb02142.x.
Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. Journal of Applied Testing Technology 14 1–38. http://jattjournal.com/index.php/atp/article/view/45254/36645.
McNamara, T. F. (1996). Measuring Second Language Performance. Harlow: Addison Wesley Longman.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 32 13–23. https://doi.org/10.3102/0013189X023002013.
Mislevy, R. J. (2012). Modeling language for assessment. In C. Chapelle (ed.), The Encyclopedia of Applied Linguistics. Hoboken, NJ: Wiley-Blackwell.
Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. Journal of Educational Measurement 53 265–292. https://doi.org/10.1111/jedm.12117.
Mislevy, R. J. (2018). Sociocognitive Foundations of Educational Measurement. London: Routledge.
Mislevy, R. J. , Behrens, J. T. , DiCerbo, K. E. and Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. Journal of Educational Data Mining 4 11–48. https://doi.org/10.5281/zenodo.3554641.
Mislevy, R. J. , Hamel, L. , Fried, R. G. , Gaffney, T. , Haertel, G. , Hafter, A. …. Wenk, A. (2003). Design Patterns for Assessing Science Inquiry (PADI Technical Report 1). SRI International. Retrieved from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf.
Mislevy, R. J. and Rahman, T. (2009). A Design Pattern for Assessing Cause and Effect Reasoning in Reading Comprehension (PADI Technical Report 20). Menlo Park, CA: SRI International.
Mislevy, R. J. , Steinberg, L. S. and Almond, R. A. (2002). Design and analysis in task-based language assessment. Language Testing 19 477–496. https://doi.org/10.1191/0265532202lt241oa.
Mislevy, R. J. , Steinberg, L. S. and Almond, R. A. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives 1 3–67. https://doi.org/10.1191/0265532202lt241oa.
Mislevy, R. J. and Yin, C. (2009). If language is a complex system, what is language assessment? Language Learning 59 (Supplement 1) 249–267. https://doi.org/10.1111/j.1467-9922.2009.00543.x.
Norris, J. M. , Brown, J. D. , Hudson, T. D. and Yoshioka, J. K. (1998). Designing Second Language Performance Assessment. Honolulu: University of Hawai'i Press.
O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. The Modern Language Journal 96 71–88. https://doi.org/10.1111/j.1540-4781.2012.01298.x.
Pearlman, M. (2008). Finalizing the test blueprint. In C. Chapelle , M. K. Enright and J. M. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language TM. London: Routledge.
Riconscente, M. M. , Mislevy, R. J. and Corrigan, S. (2015). Evidence-centered design. In S. Lane , M. R. Raymond and T. M. Haladyna (eds.), Handbook of Test Development, 2nd edn. London: Routledge.
Robinson, P. (2009). Task complexity, cognitive resources, and syllabus design. In K. V. den Branden, M. Bygate and J. M. Norris (eds.), Task-Based Language Teaching: A Reader. Amsterdam: John Benjamins.
Shute, V. J. , Masduki, I. and Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. Technology, Instruction, Cognition & Learning 8: 137–161.
Toulmin, S. E. (1958). The Uses of Argument. Cambridge: Cambridge University Press.
Weir, C. J. (ed.). (2005). Language Testing and Validation: An Evidence-Based Approach. Research and Practice in Applied Linguistics. Basingstoke: Palgrave Macmillan.
Wilson, M. (2005). Constructing Measures. Mahwah, NJ: Lawrence Erlbaum.
Young, R. F. (2008). Language and Interaction: An Advanced Resource Book. London: Routledge.

Accommodations and universal design

Abedi, J. , Zhang, Y. , Rowe, S. E. and Lee, H. (2020). Examining effectiveness and validity of accommodations for English language learners in mathematics: An evidence-based computer accommodation decision system. Educational Measurement 39 41–52. https://doi.org/10.1111/emip.12328.
Studies on the accommodations for ELLs have identified at least 73 different types of accommodations that are used by different states, many of which are adopted from the pool of accommodations created and used for students with disabilities and may not be relevant for this group of students. ELL students need assistance in the language of instruction and assessment to assimilate successfully into the mainstream

instruction and assessment; therefore, language-based accommodations would be most relevant for these students. On the other hand, not all language-based accommodations can be used for ELLs because some of these accommodations may provide unfair advantage to the recipients and can invalidate the accommodated assessment outcomes. For example, a commercial dictionary is one of the most commonly used language-based accommodations in the nation. While a dictionary can help overcome ELL students' language barriers, it provides content definitions as well, which is unfair to non-recipients of this accommodation. Based on the review of literature, we identified five different language-based accommodations that do not provide unfair advantage to the recipients and may be effective in making assessments more linguistically accessible to ELLs while at the same time not altering the focal construct (mathematics in this study). We selected mathematics as the content because the focal construct is mathematics, not language. The accommodations used in this study were (1) a linguistically modified version of the mathematics test, (2) an English glossary of non-content terms, (3) read-aloud test items, (4) a Spanish version of the test, and (5) a Spanish glossary for the test. To examine the effectiveness of the accommodations used in this study, we compared the performance of ELL students under accommodation with their performance in non-accommodated groups. Results indicated that while students under some accommodation, such as linguistically modified accommodation, consistently performed better under the accommodated assessments, the difference did not reach a statistically significant level. To examine validity of the accommodations used, we compared the performance of non-ELL students under the accommodated and non-accommodation conditions. The results of the analyses did not show any significant difference between non-ELLs who were accommodated and those tested under standard testing conditions with the original English version. These results confirm that none of the accommodations used in this study altered the focal construct.

Abedi, J. (2007). Utilizing accommodations in the assessment of English language learners. In N. H. Hornberger (ed.), Encyclopedia of Language and Education: Vol. 7: Language Testing and Assessment. Heidelberg, Germany: Springer, 331–348. Concerns over the validity of accommodations for ELL students have surfaced as the international education community moves toward inclusion of all students (including ELLs and SDs) in national and local assessments. The results of accommodated and non-accommodated assessments cannot be aggregated if the validity of accommodations has not been established. This paper summarizes the results of research on the effectiveness and validity of accommodations for ELLs and examines research findings to determine if there is sufficient evidence to inform decision makers on the provision of accommodations and reporting of accommodated assessments. While the focus of this paper is on accommodation issues for ELL students, some references have been made to accommodations for students with disabilities since ELL accommodation practices are highly affected by accommodation policies and practices for SDs.

Abedi, J. , Hofstetter, C. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based research. Review of Educational Research 74 1–28. Decisions about which accommodations to use, for whom, and under what conditions are based on limited empirical evidence of their effectiveness and validity. Given the potential consequences of test results, it is important that policymakers and educators understand the empirical base underlying their use. This article reviews test accommodation strategies for ELLs, derived from "scientifically based research." The results caution against a one-size-fits-all approach. The more promising approaches include modified English and customized dictionaries, which can be used for all students, not just ELLs.

Francis, D. , Rivera, M. , Lesaux, N. , Kieffer, M. and Rivera, H. (2006). Practical Guidelines for the Education of English Language Learners: Research-Based Recommendations for the Use of Accommodations in Large-Scale Assessments. Portsmouth, NH: RMC Research Corporation, Center on Instruction. www.centeroninstruction.org/files/ELL3-Assessments.pdf. Language is a key component of individual student success, yet ELLs are often unprepared for the rigors of academic language in content areas. This meta-analysis examines the effectiveness and validity of selected accommodations, familiar to the student through daily instructional use and applied during assessments. The results suggest the importance of linguistically appropriate accommodations, ones that have been used effectively in the classroom, and are appropriately selected to match individual student needs. Although no single accommodation has been shown to level the playing field, the most effective ones may be the use of a dictionary or glossary with extra time, provided that the student has previous classroom experience with dictionary or glossary use. Ensuring the opportunity to learn during instruction combined with accommodations that are familiar and useful to individual students in the classroom can increase the chances of academic success for all students.

Abedi, J. (2002). Assessment and accommodations of English language learners: Issues, concerns, and recommendations. Journal of School Improvement 3: 83–89.

Abedi, J. (2006). Accommodations for English Language Learners That May Alter the Construct Being Measured. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Abedi, J. (2007). English language learners with disabilities. In C. Cahlan-Laitusis and L. Cook (eds.), Accommodating Students with Disabilities on State Assessments: What Works? Arlington, VA: Council for

Exceptional Children.
Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. Educational Assessment 14 195–211. https://doi.org/10.1080/10627190903448851.
Abedi, J. (2010). Linguistic factors in the assessment of English language learners. In G. Walford , E. Tucker and M. Viswanathan (eds.), The Sage Handbook of Measurement. Oxford, UK: Sage Publications.
Abedi, J. and Ewers, N. (2013). Smarter Balanced Assessment Consortium: Accommodations for English language learners and students with disabilities: A research-based decision algorithm. Olympia, WA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Accomodations-for-under-represented-students.pdf
Abedi, J. and Herman, J. L. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. Teachers College Record 112: 723–746.
Abedi, J. , Lord, C. and Plummer, J. (1997). Language Background as a Variable in NAEP Mathematics Performance. CSE Technical Report No. 429. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Abedi, J. , Hofstetter, C. , Lord, C. and Baker, E. (1998). NAEP Math Performance and Test Accommodations: Interactions with Student Language Background. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Abedi, J. , Lord, C. , Hofstetter, C. and Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. Educational Measurement: Issues and Practice 19 16–26. https://doi.org/10.1111/j.1745-3992.2000.tb00034.x.
Abedi, J. , Courtney, M. , Mirocha, J. , Leon, S. and Goldberg, J. (2001). Language Accommodation for Large- Scale Assessment in Science. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Abedi, J. , Courtney, M. and Leon, S. (2003a). Effectiveness and Validity of Accommodations for English Language Learners in Large-Scale Assessments. CSE Technical Report No. 608. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Abedi, J. , Leon, S. and Mirocha, J. (2003b). Impact of Students' Language Background on Content-Based Assessment: Analyses of Extant Data. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Abedi, J. , Hofstetter, C. H. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. Review of Educational Research 74 1–28. https://doi.org/10.3102/00346543074001001.
Bielinski, J. , Thurlow, M. , Ysseldyke, J. , Freidebach, J. and Freidebach, M. (2001). Read-Aloud Accommodation: Effects on Multiple-Choice Reading and Math Items. Technical Report. Minneapolis, MN: National Center on Educational Outcomes.
Capp, M. J. (2017). The effectiveness of universal design for learning: A metaanalysis of literature between 2013 and 2016. International Journal of Inclusive Education 21 791–807. https://doi.org/10.1080/13603116.2017.1325074.
Chiu, C. W. T. and Pearson, D. (1999). 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, UT.
Francis, D. , Lesaux, N. , Kieffer, M. and Rivera, H. (2006). Research-Based Recommendations for the Use of Accommodations in Large-Scale Assessments. Houston, TX: Center on Instruction.
Fuchs, L. S. , Fuchs, D. , Eaton, S. B. , Hamlett, C. , Binkley, E. and Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. Exceptional Children 67 67–81. https://doi.org/10.1177/001440290006700105.
Goegan, L. D. , Radil, A. I. and Daniels, L. M. (2018). Accessibility in questionnaire research: Integrating universal design to increase the participation of individuals with learning disabilities. Learning Disabilities: A Contemporary Journal 16 177–190. https://doi.org/10.7939/r3-cmkq-1c82.
Gay, L. R. (1981). Educational Research: Competencies for Analysis and Application, 2nd edn. Columbus, OH: Charles E. Merrill.
Hafner, A. L. (2001). Evaluating the Impact of Test Accommodations on Test Scores of LEP Students & Non-LEP Students. Los Angeles, CA: California State University.
Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (ed.), Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ/: Lawrence Erlbaum, 89–116.
Islam, C. and Park, M. (2015). Preparing teachers to promote culturally relevant teaching: Helping English language learners in the classroom. Multicultural Education 23: 38–44.
Košak-Babuder, M. , Kormos, J. , Ratajczak, M. and Pižorn, K. (2019). The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. Language Testing 36 51–75. https://doi.org/10.1177/0265532218756946.

Maihoff, N. A. (2002, Junex). 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, CA.

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, UT.

Meloy, L. L. , Deville, C. and Frisbie, D. (2000, April 24–28). 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, LA. Retrieved from http://eric.ed.gov/PDFS/ED441008.pdf.

Munger, G. F. and Loyd, B. H. (1991). Effect of speediness on test performance of handicapped and nonhandicapped examinees. Journal of Educational Research 85 53–57. https://doi.org/10.1080/00220671.1991.10702812.

Munro, J. , Abbott, M. and Rossiter, M. (2013). Mentoring for success: Accommodation strategies for ELLs. Canadian Journal of Action Research 14 22–38. https://doi.org/10.33524/cjar.v14i2.83.

Rivera, C. and Collum, E. (2005). State Assessment Policies for English Language Learners: A National Perspective. New York: Routledge.

Rivera, C. and 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, WA.

Rivera, C. , Stansfield, C. W. , Scialdone, L. and Sharkey, M. (2000, April 12). An Analysis of State Policies for the Inclusion and Accommodation of English Language Learners in State Assessment Programs During 1998–99. Final Report. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA. Retrieved from http://deepwater.org/bioteacher/12-Review%20and%20SOLs/SOL-Delaware-simplification.pdf.

Roohr, K. and Sireci, S. (2017). Evaluating computer-based test accommodations for English learners. Educational Assessment 22 35–53. doi:10.1080/10627197.2016.1271704

Roohr, K. C. and Stephan, G. (2017). Evaluating computer-based test accommodations for English learners. Educational Assessment 22 35–53. https://doi.org/10.1080/10627197.2016.1271704.

Sato, E. , Rabinowitz, S. , Gallagher, C. and Huang, C. W. (2010). Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets. NCEE 2009-4079. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shaftel, J. , Belton-Kocher, E. , Glasnapp, D. and Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. Educational Assessment 11 105–126. https://doi.org/10.1207/s15326977ea1102_2.

Sireci, S. G. , Li, S. and Scarpati, S. (2003). The Effects of Test Accommodation on Test Performance: A Review of the Literature. Center for Educational Assessment Research Report No. 485. Amherst, MA: School of Education, University of Massachusetts.

Solano-Flores, G. (2012). Translation Accommodations Framework for Testing English Language Learners in Mathematics. Submitted to Smarter Balanced Assessment Consortium, University of Colorado, Boulder.

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, WA.

Thurlow, M. and Liu, K. (2001). State and District Assessments as an Avenue to Equity and Excellence for English Language Learners with Disabilities. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. House of Representatives . (2001). No Child Left Behind Act of 2001. 107th Congress 1st Session, Report 107-334. Washington, DC: Author.

Willner, L. S. , Rivera, C. and Acosta, B. D. (2008). Descriptive Study of State Assessment Policies for Accommodating English Language Learners. Arlington, VA: George Washington University Center for Equity and Excellence in Education.

Wolf, M. K. , Herman, J. L. , Kim, J. , Abedi, J. , Leon, S. , Griffin, N. …. Shin, H. W. (2008). Providing Validity Evidence to Improve the Assessment of English Language Learners. CRESST Report 738. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from https://cresst.org/publications/cresst-publication-3109/.

Wolf, M. K. , Kim, J. and Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. Applied Measurement in Education 25 347–374. https://doi.org/10.1080/08957347.2012.714693.

Rater and interlocutor training

Brown, A. (2005). Interviewer Variability in Language Proficiency Interviews. Frankfurt am Main: Peter Lang. Brown's study of a language proficiency interview documents the impact of the interviewer on scores, how interviewer behaviors influenced discourse with the test taker, and how raters compensated (or didn't) for the interviewer's performance. This study remains a model for how a variety of techniques (MFRM, conversation analysis, verbal report) can be brought to bear to understand the impact of interlocutor behavior.

Eckes, T. (2011). Introduction to Many-Facet Rasch Measurement. Frankfurt am Main: Peter Lang. This concise and accessible volume describes the use of MFRM to analyze language performance assessments. The book also describes various types of rater behavior and how such behaviors can be detected using MFRM.

Knoch, U. and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. Language Testing 35 477–499. This article provides a detailed exploration of how various aspects of the rating process impact the claims in a validity argument. Examples are provided that illustrate how a validity argument might be used to frame research into raters and scoring.

Lumley, T. (2005). Assessing Second Language Writing: The Rater's Perspective. Frankfurt am Main: Peter Lang. Lumley's book provides a detailed view of how raters in a writing assessment dealt with the challenge of interpreting scoring materials, which typically provide only minimalist descriptions of performance. Lumley also explores the usefulness and limitations of think-aloud protocols as a method for understanding rater cognition. The study provides another useful example of how quantitative and qualitative data can be combined to understand the scoring process.

American Council on the Teaching of Foreign Language (ACTFL) . (2020). ACTFL OPI Tester Certification Information Packet. ACTFL. Retrieved from www.actfl.org/sites/default/files/assessments/ACTFL%20OPI%20Tester%20Certification%20Packet%202020.pdf.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. Language Testing 33 99–115. https://doi.org/10.1177/0265532215582283.

Bachman, L. F. , Lynch, B. K. and Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. Language Testing 12 238–257. https://doi.org/10.1177/026553229501200206.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. Language Assessment Quarterly 9 225–248. https://doi.org/10.1080/15434303.2011.637262.

Ballard, L. (2017). The Effects of Primacy on Rater Cognition: An Eye Tracking Study. Doctoral dissertation. Michigan State University. MSU Electronic Theses and Dissertations. https://doi.org/10.25335/M5SR15.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. Language Assessment Quarterly 7 54–74. https://doi.org/10.1080/15434300903464418.

Bock, R. D. , Brennan, R. L. and Muraki, E. (2002). The information in multiple ratings. Applied Psychological Measurement 26 364–375. https://doi.org/10.1177/014662102237794.

Brooks, R. L. (2013). Comparing Native and Non-Native Raters of US Federal Government Speaking Tests. Doctoral dissertation. Georgetown University. Georgetown University Graduate Theses and Dissertations. Retrieved from http://hdl.handle.net/10822/559500.

Brown, A. (2005). Interviewer Variability in Language Proficiency Interviews. Frankfurt am Main: Peter Lang.

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor and P. Falvey (eds.), IELTS Collected Papers. Cambridge/: Cambridge University Press, 98–139.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London/: Routledge, 413–425.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. Language Assessment Quarterly 12 262–282. https://doi.org/10.1080/15434303.2015.1053134.

Carey, M. D. , Mannell, R. H. and Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? Language Testing 28 201–219. https://doi.org/10.1177/0265532210393704.

Choi, I. and Wolfe, E. W. (2020). The impact of operational scoring experience and additional mentored training on raters' essay scoring accuracy. Applied Measurement in Education 33 210–222. https://doi.org/10.1080/08957347.2020.1750404.

Cumming, A. , Kantor, R. and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. The Modern Language Journal 86 67–96. https://doi.org/10.1111/1540-4781.00137.

Davis, L. E. (2012). Rater Expertise in a Second Language Speaking Assessment: The Influence of Training and Experience. Doctoral dissertation. University of Hawai'i, Manoa. ScholarSpace, University of Hawaii at Manoa. Retrieved from https://scholarspace.manoa.hawaii.edu/handle/10125/100897.

Davis, L. E. (2016). The influence of training and experience on rater performance in scoring spoken language. Language Testing 33 117–135. https://doi.org/10.1177/0265532215582282.

Davis, L. E. (2019). Rater training in a speaking assessment: Impact on more- and less-proficient raters. In S. Papageorgiou and K. Bailey (eds.), Global Perspectives on Language Assessment. London/: Routledge, 18–31.
Diederich, P. B. , French, J. W. and Carlton, S. T. (1961). Factors in Judgements of Writing Ability. Research Bulletin, RB-61-15. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x.
Douglas, D. and Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida. Cambridge/: Cambridge University Press, 60–81.
Duijm, K. , Schoonen, R. and Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. Language Testing 35 501–527. https://doi.org/10.1177/0265532217712553.
Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. Language Assessment Quarterly 2 197–221. https://doi.org/10.1207/s15434311laq0203_2.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. Language Testing 25 155–185. https://doi.org/10.1177/0265532207086780.
Eckes, T. (2011). Introduction to Many-Facet Rasch Measurement. Frankfurt am Main: Peter Lang.
Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. Language Assessment Quarterly 9 270–292. https://doi.org/10.1080/15434303.2011.649381.
Edgeworth, F. Y. (1890). The element of chance in competitive examinations. Journal of the Royal Statistical Society 53 644–663. Retrieved from www.jstor.org/stable/2979547.
Elder, C. , Barkhuizen, G. , Knoch, U. and Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. Language Testing 24 37–64. https://doi.org/10.1177/0265532207071511.
Ericsson, K. A. (2006). An introduction to the Cambridge handbook of expertise and expert performance: Its development, organization, and content. In K. A. Ericsson , N. Charness , P. J. Feltovich and R. R. Hoffman (eds.), The Cambridge Handbook of Expertise and Expert Performance. Cambridge/: Cambridge University Press, 3–19.
Ericsson, K. A. and Simon, H. A. (1993). Protocol Analysis: Verbal Reports as Data, 2nd edn. Cambridge, MA: MIT Press.
Everson, P. and Hines, S. (2010). How ETS scores the TOEIC® speaking and writing test responses. In D. E. Powers (ed.), The Research Foundation for the TOEIC® Tests: A Compendium of Studies, vol. II. Princeton, NJ: Educational Testing Service, 8.1–8.9.
Fahim, M. and Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. Iranian Journal of Language Testing 1: 1–16.
Feltovich, P. J. , Prietula, M. J. and Ericsson, K. A. (2006). Studies of expertise from psychological perspectives. In K. A. Ericsson , N. Charness , P. J. Feltovich and R. R. Hoffman (eds.), The Cambridge Handbook of Expertise and Expert Performance. Cambridge/: Cambridge University Press, 41–67.
Fulcher, G. (2003). Testing second language speaking. Harlow: Longman.
Furneaux, C. and Rignall, M. (2007). The effect of standardization – training on rater judgements for the IELTS writing module. In L. Taylor and P. Falvey (eds.), IELTS Collected Papers: Research in Speaking and Writing Assessment. Cambridge/: Cambridge University Press, 422–445.
Hamp-Lyons, L. (2007). Worrying about rating. Assessing Writing 1 1–9. https://doi.org/10.1016/j.asw.2007.05.002.
Hsieh, C. N. (2011). Rater Effects in ITA Testing: ESL Teachers' Versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency. Doctoral dissertation. Michigan State University. MSU Electronic Theses and Dissertations. Retrieved from https://d.lib.msu.edu/etd/1282.
Isaacs, T. and Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. Language Assessment Quarterly 10 135–159. https://doi.org/10.1080/15434303.2013.769545.
Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. Language Assessment Quarterly 9 249–269. https://doi.org/10.1080/15434303.2011.642631.
Kim, A. Y. and Gennaro, K. D. (2012). Scoring behavior of native vs. non-native speaker raters of writing exams. Language Research 48 319–342. Retrieved from http://hdl.handle.net/10371/86486.
Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. Language Assessment Quarterly 12 239–261. https://doi.org/10.1080/15434303.2015.1049353.
Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. Language Testing 26 187–217. https://doi.org/10.1177/0265532208101010.

Knoch, U. (2009). Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale. Spaan Fellow Working Papers in Second or Foreign Language Assessment 7: 21–46.
Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. Language Testing 28 179–200. https://doi.org/10.1177/0265532210384252.
Knoch, U. and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. Language Testing 35 477–499. https://doi.org/10.1177/0265532217710049.
Knoch, U. , Fairbairn, J. , Myford, C. and Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. Papers in Language Testing and Assessment 7: 61–86.
Knoch, U. , Read, J. and von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? Assessing writing 12 26–43. https://doi.org/10.1016/j.asw.2007.04.001.
Kuiken, F. and Vedder, I. (2014). Rating written performance: What do raters do and why? Language Testing 31 329–348. https://doi.org/10.1177/0265532214526174.
Lave, J. and Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge: Cambridge University Press.
Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. Language Testing 13 151–172. https://doi.org/10.1177/026553229601300202.
Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. Language Testing 28 543–560. https://doi.org/10.1177/0265532211406422.
Lumley, T. (2005). Assessing Second Language Writing: The Rater's Perspective. Frankfurt: Peter Lang.
Lynch, B. K. and McNamara, 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 15 158–180. https://doi.org/10.1177/026553229801500202.
Mackey, A. and Gass, S. M. (2016). Stimulated Recall Methodology in Applied Linguistics and L2 Research. London: Routledge.
Milanovic, M. , Saville, N. and Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic and N. Saville (eds.), Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem. Cambridge/: Cambridge University Press, 92–114.
Mislevy, R. J. (2010). Some implications of expertise research for educational assessment. Research Papers in Education 25 253–270. https://doi.org/10.1080/02671522.2010.498142.
Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher 23 5–12. https://doi.org/10.3102/0013189X023002005.
Myford, C. M. and Wolfe, E. W. (2000). Monitoring Sources of Variability Within the Test of Spoken English Assessment System. ETS Research Report Series. Report No. RR-00-6. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x.
Nakatsuhara, F. , Inoue, C. and Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. Language Assessment Quarterly. Advance online publication. https://doi.org/10.1080/15434303.2020.1799222.
O'Hagan, S. R. and Wigglesworth, G. (2015). Who's marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context. Studies in Higher Education 40 1729–1747. https://doi.org/10.1080/03075079.2014.896890.
Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. System 30 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7.
O'Sullivan, B. and Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor and P. Falvey (eds.), IELTS Collected Papers: Research in Speaking and Writing Assessment. Cambridge/: Cambridge University Press, 446–478.
Papajohn, D. (2002). Concept mapping for rater training. TESOL Quarterly 36 219–233. https://doi.org/10.2307/3588333.
Randall, J. and Engelhard Jr, G. (2009). Examining teacher grades using Rasch measurement theory. Journal of Educational Measurement 46 1–18. https://doi.org/10.1111/j.1745-3984.2009.01066.x.
Ross, S. (1992). Accommodative questions in oral proficiency interviews. Language Testing 9 173–185. https://doi.org/10.1177/026553229200900205.
Şahan, Ö. and Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? Language Testing 37 311–332. https://doi.org/10.1177/0265532219900228.
Saito, K. , Trofimovich, P. , Isaacs, T. and Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs and P. Trofimovich (eds.), Second Language Pronunciation Assessment: Interdisciplinary Perspectives. Multilingual Matters, 141–156. Retrieved from www.jstor.org/stable/10.21832/j.ctt1xp3wcc.12.
Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright and F. Bolger (eds.), Expertise and Decision Support. New York/: Plenum, 11–23.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. Research Notes 8 13–17. www.cambridgeenglish.org/Images/23120-research-notes-08.pdf.
Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ/: Educational Testing Service, 1–12.
Spolsky, B. (1995). Measured Words. Oxford: Oxford University Press.
Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), Examining Speaking: Studies in Language Testing 30. Cambridge/: Cambridge University Press, 171–233.
Wei, J. and Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. Language Assessment Quarterly 12 283–304. https://doi.org/10.1080/15434303.2015.1037446.
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. Language Testing 11 197–223. https://doi.org/10.1177/026553229401100206.
Weigle, S. C. (1998). Using FACETS to model rater training effects. Language testing 15 263–287. https://doi.org/10.1177/026553229801500205.
Weigle, S. C. (2002). Assessing Writing. Cambridge: Cambridge University Press.
Weir, C. J. (2003). A survey of the history of the certificate of proficiency in English (CPE) in the twentieth century. In C. J. Weir and M. Milanovic (eds.), Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002. Cambridge/: Cambridge University Press, 1–56.
Weir, C. J. (2005). Language Testing and Validation. Houndgrave: Palgrave Macmillan.
White, E. M. (1984). Holisticism. College Composition and Communication 35 400–409. https://doi.org/10.2307/357792.
White, E. M. (1985). Teaching and Assessing Writing. San Francisco: Jossey-Bass.
Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. Language Testing 10 305–319. https://doi.org/10.1177/026553229301000306.
Wilds, C. P. (1975). The oral interview test. In R. L. Jones and B. Spolsky (eds.), Testing Language Proficiency. Arlington, VA/: Center for Applied Linguistics, 29–44.
Winke, P. , Gass, S. and Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. Language Testing 30 231–252. https://doi.org/10.1177/0265532212456968.
Winke, P. and Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. Assessing Writing 25 37–53. https://doi.org/10.1016/j.asw.2015.05.002.
Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. Journal of Writing Assessment 2: 37–56.
Wolfe, E. W. , Matthews, S. and Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. Journal of Technology, Learning, and Assessment 10. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1601.
Wolfe, E. W. , Song, T. and Jiao, H. (2016). Features of difficult-to-score essays. Assessing Writing 27 1–10. https://doi.org/10.1016/j.asw.2015.06.002.
Xi, X. and Mollaun, P. (2009). How Do Raters from India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?. Research Report No. RR-09-31. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x.
Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. Assessing Writing 27 37–53. https://doi.org/10.1016/j.asw.2015.11.001.
Zhang, Y. and Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? Language Testing 28 31–50. https://doi.org/10.1177/0265532209360671.
Zhang, Y. and Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the college English test-spoken English test (CET-SET). Assessment in Education: Principles, Policy & Practice 21 306–325. https://doi.org/10.1080/0969594X.2013.845547.

Item writing and item writers

Alderson, J. C. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge, UK: Cambridge University Press. Chapters 2 and 3 of this volume have useful information on test specs, item types, and item writing. Their practical tips include (1) specs need to be adjusted to different audiences for different contexts and (2) item writing is best practiced in groups. Their guidelines remain effective for and relevant to recent discussions of the item-writing process.

Bachman, L. F. and Palmer, A. S. (2010). Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World. Oxford, UK: Oxford University Press. This book's strength is that it elaborates all the nuts and bolts of the assessment use argument (AUA) framework for the entire test development and validation process. The authors have identified a great deal of useful information by conceptualizing test design and score interpretation through AUA and by illustrating specific examples for different contexts.

Davidson, F. and Lynch, B. K. (2002). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. New Haven, CT: Yale University Press. This concisely written book focuses on applying test specs to the creation of valid test items. Item writers' experience and wisdom, in the forms of principles or guidelines, are integrated to create a unified system for the specs-to-item writing process. This idea has been expanded in a recent publication on test retrofit (Fulcher and Davidson, 2009), much of which is metaphorically explained using architecture.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York, NY: Routledge. The practice of item development requires analytic and integrated language testing skills. This book is well designed to teach practitioners these skills and includes extensive discussion and application exercises. The notion of effect-driven testing is useful for understanding how item writing, as an iterative process, needs to be transparently regarded within the full picture of test development and validation.

Shohamy, E. (2001). The Power of Tests: A Critical Perspective on the Uses and Consequences of Language Tests. London, UK: Longman. Although this book may appear disconnected from work containing guidelines/systems for item development, it effectively explains why item writing should not be viewed only as a science. Stohamy's claims that intended effects can lead to unintended and unexpected consequences deserve thoughtful attention in all areas of test design.

Alderson, J. C. (2000). Assessing Reading. Cambridge, UK: Cambridge University Press.

Alderson, J. C. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. Language Testing 17 1–42. https://doi.org/10.1177/026553220001700101.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice. Oxford, UK: Oxford University Press.

Brown, J. D. and Hudson, T. D. (1998). The alternatives in language assessment. TESOL Quarterly 32 653–675. https://doi.org/10.2307/3587999.

Buck, G. (2001). Assessing Listening. Cambridge, UK: Cambridge University Press.

Chalhoub-Deville, M. (ed.). (1999). Issues in Computer-Adaptive Testing of Reading Proficiency. Cambridge, UK: Cambridge University Press.

Constantinou, F. , Crisp, V. and Johnson, M. (2018). Multiple voices in tests: Toward a macro theory of test writing. Cambridge Journal of Education 48 411–426. https://doi.org/10.1080/0305764X.2017.1337723.

Cronbach, L. J. (1970). Book review. [Review of the book on The Theory of Achievement Test Items, by J. R. Bormuth] Psychometrika 35: 509–511.

Davidson, F. and Lynch, B. K. (2002). Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications. New Haven, CT: Yale University Press.

Douglas, D. (2000). Assessing Language for Specific Purposes. Cambridge, UK: Cambridge University Press.

Downing, S. M. and Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. Applied Measurement in Education 10 61–82. https://doi.org/10.1207/s15324818ame1001_4.

Downing, S. M. and Haladyna, T. M. (eds.). (2006). Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.

Elwood, J. and Murphy, P. (2015). Assessment systems as cultural scripts: A sociocultural theoretical lens on assessment practice and products. Assessment in Education: Principles, Policy & Practice 22 182–192. https://doi.org/10.1080/0969594X.2015.1021568.

Fulcher, G. (2000). The "communicative" legacy in language testing. System 28 483–497. https://doi.org/10.1016/S0346-251X(00)00033-6.

Fulcher, G. (2003). Testing Second Language Speaking. London, UK: Longman.

Fulcher, G. (2009). Test use and political philosophy. Annual Review of Applied Linguistics 29 3–20. https://doi.org/10.1017/S0267190509090023.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London, UK: Routledge.

Fulcher, G. and Davidson, F. (2009). Test architecture: Test retrofit. Language Testing 26 123–144. https://doi.org/10.1177/0265532208097339.

Green, A. and Hawkey, R. (2011). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. Language Testing 29 109–129.

https://doi.org/10.1177/0265532211413445.
Haladyna, T. M. and Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item writing rules. Applied Measurement in Education 1 51–78. https://doi.org/10.1207/s15324818ame0201_4.
Haladyna, T. M. , Downing, S. M. and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education 15 309–334. https://doi.org/10.1207/S15324818AME1503_5.
Hamp-Lyons, L. (1997). Ethics in language testing. In C. Clapham and D. Corson (eds.), Encyclopedia of Language and Education, Vol. 7. Language Testing and Assessment. Dordrecht, Netherlands/: Kluwer Academic Publishers, 323–333.
Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. System 28 579–591. https://doi.org/10.1016/S0346-251X(00)00039-7.
Hamp-Lyons, L. and Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (ed.), Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach. Mahwah, NJ/: Lawrence Erlbaum, 253–276.
Harsch, C. (2018). Developing test specifications for listening. In J. I. Liontas , T. International Association and M. DelliCarpini (eds.), The TESOL Encyclopedia of English Language Teaching. https://doi.org/10.1002/9781118784235.eelt0619.
Hughes, A. (2003). Testing for Language Teachers. Cambridge, UK: Cambridge University Press.
Ingram, K. (2008). The Cambridge ESOL approach to item writer training: The case of ICFE listening. Research Note 32: 5–9.
Johnson, M. , Constantinou, F. and Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. British Educational Research Journal 43 700–719. https://doi.org/10.1002/berj.3281.
Kim, J. (2008). Development and Validation of an ESL Diagnostic Reading-to-Write Test: An Effect-Driven Approach. PhD thesis, University of Illinois at Urbana-Champaign.
Kim, J. , Chi, Y. , Huensch, A. , Jun, H. , Li, H. and Roullion, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writer's characteristics. Language Assessment Quarterly 7 160–174. https://doi.org/10.1080/15434300903473989.
Kohn, A. (2000). The Case Against Standardized Testing: Raising the Scores, Ruining the Schools. Portsmouth, NH: Heinemann.
Luoma, S. (2004). Assessing Speaking. Cambridge, UK: Cambridge University Press.
Lynch, B. K. (2001). Rethinking assessment from a critical perspective. Language Testing 18 351–372. https://doi.org/10.1177/026553220101800403.
Lynch, B. K. and Shaw, P. (2005). Portfolio, power, and ethics. TESOL Quarterly 39 263–297. https://doi.org/10.2307/3588311.
McNamara, T. (2000). Language Testing. Oxford, UK: Oxford University Press.
McNamara, T. (2008). The socio-political and power dimensions of tests. In E. Shohamy and N. H. Hornberger (eds.), Encyclopedia of Language and Education, Vol 7. Language Testing and Assessment, 2nd edn. New York, NY: Springer, 415–427.
McNamara, T. and Roever, C. (2006). Language Testing: The Social Dimension. London, UK: Blackwell Publishing.
Mislevy, R. J. , Steinberg, L. S. and Almond, R. G. (2002). Design and analysis in task-based language assessment. Language Testing 19, 477–496. https://doi.org/10.1191/0265532202lt241oa.
Mislevy, R. J. , Steinberg, L. S. and Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives 1 3–62. https://doi.org/10.1207/S15366359MEA0101_02.
Morgan, B. (2007). Poststructuralism and applied linguistics: Complementary approaches to identity and culture in ELT. In J. Cummins and C. Davison (eds.), Springer International Handbook of Education, Vol. 15. International Handbook of English Language Teaching. Norwell, MA/: Springer, 949–968.
Ngo, X. M. (2016, October). Demystifying Item Writing: The Need for a Theoretical Framework. Paper presented at the 4th British Council New Directions in English Language Assessment, Conference, Hanoi, Vietnam.
Norton, B. (2000). Identity and Language Learning: Gender, Ethnicity, and Educational Change. London, UK: Longman.
Oller, J. W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. Modern Language Journal 56 151–158. https://doi.org/10.2307/324037.
Peirce, B. N. (1992). Demystifying the TOEFL reading test. TESOL Quarterly 26 665–691. https://doi.org/10.2307/3586868.
Pennycook, A. (2001). Critical Applied Linguistics: A Critical Approach. Mahwah, NJ: Lawrence Erlbaum.
Pugh, D. , De Champlain, A. , Gierl, M. , Lai, H. and Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? RPTEL 15.

https://doi.org/10.1186/s41039-020-00134-8.
Purpura, J. E. (2004). Assessing Grammar. Cambridge, UK: Cambridge University Press.
Read, J. (2000). Assessing Vocabulary. Cambridge, UK: Cambridge University Press.
Roid, G. H. and Haladyna, T. M. (1982). A Technology for Test-item Writing. New York, NY: Academic Press.
Ryan, È. and Brunfaut, T. (2016). When the test developer does not speak the target language: The use of language informants in the test development process. Language Assessment Quarterly 13 393–408. https://doi.org/10.1080/15434303.2016.1236110.
Salisbury, K. (2005). The Edge of Expertise: Towards an Understanding of Listening Test Item Writing as Professional Practice. PhD thesis, Kings College London.
Shin, D. and Kim, N. (2010). Munhang Kaebal Kwajungeseo Shihumjaksung Sepukyehwukseo Yuk-whalyeonKu [Investigating the role of test specifications in item development process]. Foreign Languages Education 17: 257–279.
Shizuka, T. , Takeuchi, O. , Yashima, T. and Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. Language Testing 23 35–57. https://doi.org/10.1191/0265532206lt319oa.
Shohamy, E. (2001). The Power of Tests: A Critical Perspective on the Uses and Consequences of Language Tests. London, UK: Longman.
Spaan, M. (2007). Evolution of a test item. Language Assessment Quarterly 4 279–293. https://doi.org/10.1080/15434300701462937.
Spaan, M. (2006). Test and item specifications development. Language Assessment Quarterly 3 71–79. https://doi.org/10.1207/s15434311laq0301_5.
Spolsky, B. (1995). Measured Words. Oxford, UK: Oxford University Press.
Stubbs, J. B. and Tucker, G. R. (1974). The cloze test as a measure of English proficiency. Modern Language Journal 58 239–241. https://doi.org/10.2307/325020.
Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. Language Assessment Quarterly 8 34–52. https://doi.org/10.1080/15434303.2010.536924.
Tunks, J. (2001). The effect of training in test item writing on test performance of junior high students. Educational Studies 27 129–142. https://doi.org/10.1080/03055690120050374.
Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language test specification reverse engineering. Language Assessment Quarterly 10 317–342. https://doi.org/10.1080/15434303.2010.516042.
Weigle, S. W. (2002). Assessing Writing. Cambridge, UK: Cambridge University Press.
Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? Language Testing 24 539–572. https://doi.org/10.1177/0265532207080780.
Zandi, H. , Kaivanpanah, S. and Alavi, S. (2014). The effect of test specifications review on improving the quality of a test. Iranian Journal of Language Teaching Research 2: 1–14.

Writing integrated tasks

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability study. Assessing Writing 15: 100–117. Using multivariate generalizability analysis, Gebril investigates reliability between independent and integrated tasks by comparing raters and task facets with the two tasks combined and with each task separately. His findings uncover similar results in reliability for the two combined and as separate tests, which provides evidence that the tasks are measuring the same construct. The results also exhibit similar results when different raters scored the two tasks and when the same raters scored the tasks. Gebril advocates using independent and integrated tasks in combination when assessing writing to provide comprehensive information on second language writing ability.
Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. Language Assessment Quarterly 10: 1–8. This editorial introduced a special issue on integrated writing task for assessment anchored around the "promises" and "perils" he predicts as a leading scholar in the field. The promises that he sees in these tasks, which emerged in the special issue, are that the tasks (1) elicit realistic and complex literacy engagement; (2) require test takers to draw on sources; (3) are performance based, potentially avoiding a method effect and providing diagnostic or instructional use; and (4) align with construction-integration or multiliteracies views of literacy. The perils substantiated in the research of the special issue are also detailed. The construct issue emerges, which confounds writing with reading and skills required in using source materials. These tasks also are seen to require a threshold level of ability in English and may not fall on a clearly linear progression of ability levels. In addition, he finds the genres for authentic source-based writing are not well defined, especially in terms of scoring. Related to the practical issues of

scoring, integrated writing can elicit texts that blur the distinction between a test taker's writing and that of the source materials. His concluding thoughts are that integrated tasks should continue to be developed and provide opportunities to assess literacy in more depth than independent writing; however, ongoing research is necessary to continue with the complexity introduced by this multi-skill, multi-text writing task.

Wang, J. , Engelhard, G. , Raczynski, K. , Song, T. and Wolfe, E. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach, Assessing Writing 33: 36–47. This study focused on source-based writing and rating using a mixed-methods design, which first quantitatively identified difficult-to-score essays, then qualitatively analyzed survey data from raters to explain the difficulty. There were four essay features mentioned consistently by raters when marking a difficult essay: (1) focus of the essay, (2) textual borrowing from source materials, (3) original development of ideas, and (4) essay organization. For these features in hard-to-score essays, raters and experts showed a lack of agreement and overall inconsistency when scoring. The study implications advocate care in preparing rater training materials to include effective prototype essays for training and selective materials to emphasize features that make rating difficult. Furthermore, rater retraining, especially with changes in source materials, is seen as important to maintain reliable scoring for integrated tasks.

Ascención, Y. (2005). Validation of Reading-to-write Assessment Tasks Performed by Second Language Learners. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. Journal of Second Language Writing 18 191–208. https://doi.org/10.1016/j.jslw.2009.05.003.

Brown, J. D. , Hilgers, T. and Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. Written Communication 8 533–556. https://doi.org/10.1177/0741088391008004005.

Carroll, J. M. (1961). Fundamental Considerations in Teaching for English Language Proficiency of Foreign Students. Washington, DC: Routledge.

Chalhoub-Deville, M. and Deville, C. (2005). Looking back at and forward to what language testers measure. In E. Hinkel (ed.), Handbook of Research on Second Language Teaching and Learning. Mahwah, NJ: Erlbaum.

Cho, Y. and Choi, I. (2018). Writing from sources: Does audience matter? Assessing Writing 37 25–38. https://doi.org/10.1016/j.asw.2018.03.004.

Cho, Y. , Rijmen, F. and Novák, J. (2013). Investigating the effect of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. Language Testing 30 513–534. https://doi.org/10.1177/0265532213478796.

Cooper, R. L. (1965). Testing. In H. B. Allen and R. N. Allen (eds.), Teaching English as a Second Language: A Book of Readings. New York, NY: McGraw-Hill.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. Language Assessment Quarterly 10 1–8. https://doi.org/10.1080/15434303.2011.622016.

Cumming, A. , Kantor, R. , Baba, K. , Erdosy, U. , Eouanzoui, K. and James, M. (2005). Differences in written discourse in writing-only and reading-to-write prototype tasks for next generation TOEFL. Assessing Writing 10: 5–43.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. Journal of English for Academic Purposes 7 140–150. https://doi.org/10.1016/j.jeap.2008.04.001.

Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. Canadian Modern Language Journal 58 599–622. https://doi.org/10.3138/cmlr.58.4.599.

Farhady, H. (1979). The disjunctive fallacy between discrete point and integrative tests. TESOL Quarterly 13 347–357. https://doi.org/10.2307/3585882.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (ed.), Issues in Language Testing Research. Rowley, MA: Newbury House Publishers.

Fitzgrerald, J. and Shanahan, T. (2000). Reading and writing relations and their development. Educational Psychologist 35 39–50. https://doi.org/10.1207/S15326985EP3501_5.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? Language Testing 26 507–531. https://doi.org/10.1177/0265532209340188.

Gebril, A. and Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. Spaan Working Papers in Second or Foreign Language Assessment 7: 47–84.

Gebril, A. and Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. Assessing Writing, 21 56–73. https://doi.org/10.1016/j.asw.2014.03.002.

Grabe, W. and Cui, Z. (2016). Reading-writing relationships in first and second language academic literacy development. Language Teaching 49 339–355. https://doi.org/10.1017/S0261444816000082.

Huang, H. T. and Hung, S. T. (2013). Comparing the effects of test anxiety on independent and integrated speaking performance. TESOL Quarterly 47 : 444–49. https://doi.org/10.1002/tesq.69.

Huang, H. T. , Huang, S. T. and Plakans, L. (2018). Topical knowledge and L2 speaking assessment: Comparing independent and integrated speaking assessments. Language Testing 35 27–49. https://doi.org/10.1177/0265532216677106.

Knoch, U. and Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. Assessing Writing 18 300–308. https://doi.org/10.1016/j.asw.2013.09.003.
Koda, K. (2004). Insights into Second Language Reading: A Cross-Linguistic Approach. Cambridge, UK: Cambridge University Press.
Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. Language Testing 23 131–166. https://doi.org/10.1191/0265532206lt325oa.
Lee, H. and Anderson, C. (2007). Validity and topic generality of a writing performance test. Language Testing 24 307–330. https://doi.org/10.1177/0265532207077200.
Leki, I. and Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. TESOL Quarterly 31 39–69. https://doi.org/10.2307/3587974.
Leki, I. , Cumming, A. and Silva, T. (2008). A Synthesis of Research on Second Language Writing in English. New York, NY: Routledge.
Lewkowicz, J. A. (1994). Writing from Sources: Does source material help or hinder students' performance? In N. Bird et al. (eds.), Language and Learning: Papers Presented at the Annual International Language in Education Conference. Hong Kong: ERIC Document (ED 386 050).
Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham and D. Corson (eds.), Encyclopaedia of Language and Education. Dordrecht, The Netherlands: Kluwer.
Li, J. (2014). Examining genre effects on test takers' summary writing performance. Assessing Writing 22 75–90. https://doi.org/10.1016/j.asw.2014.08.003.
Ohta, R. , Plakans, L. and Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scoring: A generalizability analysis. Assessing Writing 38 21–36. https://doi.org/10.1016/j.asw.2018.08.001.
Oller, J. W. (1979). Language Tests at School: A Pragmatic Approach. London, UK: Longman.
Oller, J. W. (1983). Issues in Language Testing Research. Rowley, MA: Newbury House.
Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. Assessing Writing 13 111–129. https://doi.org/10.1016/j.asw.2008.07.001.
Plakans, L. (2009a). The role of reading strategies in L2 writing tasks. Journal of English for Academic Purposes 8 252–266. https://doi.org/10.1016/j.jeap.2009.05.001.
Plakans, L. (2009b). Discourse synthesis in integrated second language assessment. Language Testing 26 1–27. https://doi.org/10.1177/0265532209340192.
Plakans, L. (2010). Independent vs. Integrated tasks: A comparison of task representation. TESOL Quarterly 44 185–194. www.jstor.org/stable/27785076.
Plakans, L. and Gebril, A. (2012). A close investigation into source use in integrated writing tasks. Assessing Writing 17 18–34. https://doi.org/10.1016/j.asw.2011.09.002.
Plakans, L. and Gebril, A. (2013). Using multiple sources in a listening and reading-based writing assessment task: Source text use as a predictor of score. Second Language Writing Journal 22 : 217–230. https://doi.org/10.1016/j.jslw.2013.02.003.
Plakans, L. , Gebril, A. and Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. Language Testing 36: 61–79. https://doi.org/10.1177/0265532216669537.
Plakans, L. , Liao, J. T. and Wang, F. (2018). Integrated assessment research: Writing-into-reading. Language Teaching 51 430–434. https://doi.org/10.1016/j.asw.2019.03.003.
Plakans, L. , Liao, J. T. and Wang, F. (2019). "I should summarize this whole paragraph": Shared processes of reading and writing in iterative integrated tasks. Assessing Writing 40: 14–26.
Read, J. (1990). Providing relevant content in an EAP writing test. English for Specific Purposes 9 109–121. https://doi.org/10.1016/0889-4906(90)90002-T.
Ruiz-Funes, M. (2001). Task representation in foreign language reading-to-write. Foreign Language Annals 34 226–234. https://doi.org/10.1111/j.1944-9720.2001.tb02404.x.
Sawaki, Y. , Stricker, L. J. and Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. Language Testing 26 5–30. https://doi.org/10.1177/0265532208097335.
Sawaki, Y. , Quinlan, T. and Lee, Y. (2013). Understanding learning strengths and weaknesses: Assessing student performance on an integrated writing task. Language Assessment Quarterly 10: 73–95. https://doi.org/10.1080/15434303.2011.633305.
Shanahan, T. and Lomax, R. (1986). An analysis and comparison of theoretical models of the reading-writing relationship. Journal of Educational Psychology 78 116–123. https://doi.org/10.1037/0022-0663.78.2.116.
Spivey, N. (1990). Transforming texts: Constructive processes in reading and writing. Written Communication 7 256–287. https://doi.org/10.1177/0741088390007002004.
Swain, M. , Huang, L. , Barkaoui, K. , Brooks, L. and Lapkin, S. (2009). The speaking section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors. TOEFL iBT Research Report, RR 09–30. Retrieved from www.ets.org/Media/Research/pdf/RR-09-30.pdf.

Trites, L. and McGroaty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? Language Testing 22 174–210. https://doi.org/10.1191/0265532205lt299oa.
Wang, J. , Engelhard, G. , Raczynski, K. , Song, T. and Wolfe, E. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. Assessing Writing 33 36–47. https://doi.org/10.1016/j.asw.2017.03.003.
Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. Assessing Writing 9 27–55. https://doi.org/10.1016/j.asw.2004.01.002.
Wolfersberger, M. A. (2007). Second Language Writing from Sources: An Ethnographic Study of an Argument Essay Task. Unpublished doctoral dissertation, The University of Auckland, Auckland, New Zealand.
Yang, H. C. (2014). Toward a model of strategies and summary writing performance. Language Assessment Quarterly 11 403–431. https://doi.org/10.1080/15434303.2014.957381.
Yang, H. C. and Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. TESOL Quarterly 46 80–103. https://doi.org/10.1002/tesq.6.
Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. Assessing Writing 14 116–137. https://doi.org/10.1016/j.asw.2009.04.002.

Test-taking strategies and task design

Bowles, M. A. (2010). The Think-Aloud Controversy in Second Language Research. Abingdon, UK: Routledge. The book deals with the validity and use of respondents' verbal reports during the performance of language tasks. After presenting theoretical background and empirical research on the validity of think alouds, the author gives an overview of how think alouds have been used in L2 language research, as well as a meta-analysis of findings from studies involving think alouds on verbal tasks. The volume also offers guidance regarding the practical issues of data collection and analysis.
Brown, J. D. (ed.). (2013). New Ways of Classroom Assessment. Revised. ERIC ED549559 (396 pp.). Alexandria, VA: Teachers of English to Speakers of Other Languages. While not focused on test-taking strategies per se, the volume constitutes a compendium of everyday classroom assessment activities that provide a way of observing or scoring students' performances and giving feedback that is meant to inform students and teachers about the effectiveness of the teaching and learning taking place. Each activity comes with suggestions as to how to give feedback in the form of a score or other information (e.g., notes in the margin, written prose reactions, oral critiques, teacher conferences). Many of the entries utilize other possible feedback perspectives aside from that of the teacher: namely, self-assessment, peer assessment, and outsider assessment – often used in conjunction with teacher assessment. One entry on "Preparing students for tests" by Alastair Allan (pp. 205–09) expressly deals with what I would term test-deviousness strategies. Although not necessarily calling upon the teacher to be a collaborator in the assessment process as does dynamic assessment (Poehner, 2007, 2008) the assessment activities in this volume are more aligned with DA than are traditional language assessment activities. For table of contents, go to www.tesol.org/read-and-publish/bookstore/toc/NW_classroomassessmentrevised_TOC (accessed April 14, 2019).
Green, A. J. F. (1998). Using Verbal Protocols in Language Testing Research: A Handbook. Cambridge, UK: Cambridge University Press. Referring to verbal report as verbal protocol analysis (VPA), the author notes that it is a complex methodology, warranting orientation of users to maximize its benefits. The book provides background regarding VPA in research on language assessment (the design of data collection methods, the development of coding schemes for analyzing the data, and the actual analysis of the data).
Gass, S. M. and Mackey, A. (2000). Simulated Recall Methodology in Second Language Research. Mahwah, NJ: Lawrence Erlbaum. The book focuses on retrospective verbal report data, referred to as *stimulated recall*, and gives recommendations for how to collect and analyze such data. The authors also consider issues of reliability and validity and uses for stimulated recall – for example, in comprehending and producing oral language; understanding the dynamics of L2 classroom interaction; looking at processes in L2 reading; and investigating L2 syntactic development, vocabulary acquisition, and pragmatics.
Kelly, M. (2019). 10 Reading Comprehension Strategies All Students Need. Retrieved April 14, 2019, from www.thoughtco.com/reading-comprehension-strategies-7952. This is just one recent offering from ThoughtCo, an internet-based company that provides many articles on language assessment, accessible through a search engine.
Ahmadian, M. , Poulaki, S. and Farahani, E. (2016). Reading strategies used by high scoring and low scoring IELTS candidates: A think-aloud study. Theory and Practice in Language Studies 6 408–416. http://dx.doi.org/10.17507/tpls.0602.25.
Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test-takers. Language Testing 9 101–122. https://doi.org/10.1177/026553229200900201.

Amer, A. A. (1993). Teaching EFL students to use a test-taking strategy. Language Testing 10 71–77. https://doi.org/10.1177/026553229301000104.
Assiri, M. (2014). Metacognitive and cognitive strategies use and performance on a reading test with multiple-format tasks. Arab World English Journal 5: 187–202.
Barkaoui, K. , Brooks, L. , Swain, M. and Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. Applied Linguistics 34 304–324. https://doi.org/10.1093/applin/ams046.
Bulushi, A. , Al Seyabi, F. and Al-Busaidi, S. (2019). The role of test taking strategies in improving Omani students' listening comprehension. International Journal of Education 10 57–77. https://doi.org/10.5296/ije.v10i4.13951.
Chang, A. C.-S. (2009). EFL listeners' task-based strategies and their relationship with listening performance. TESL-EJ 13. Retrieved April 9, 2019, from www.tesl-ej.org/wordpress/issues/volume13/ej50/ej50a1/.
Cohen, A. D. (1984). On taking language tests: What the students report. Language Testing 1 70–81. https://doi.org/10.1177/026553228400100106.
Cohen, A. D. (1993). The role of instructions in testing summarizing ability. In D. Douglas and C. Chapelle (eds.), A New Decade of Language Testing: Collaboration and Cooperation. Alexandria, VA/: TESOL, 132–160.
Cohen, A. D. (2006). The coming of age of research on test-taking strategies: Collaboration and cooperation. Language Assessment Quarterly 3 307–331. https://doi.org/10.1080/15434300701333129.
Cohen, A. D. (2007). Coming to terms with language learner strategies: Surveying the experts. In A. D. Cohen and E. Macaro (eds.), Language Learner Strategies: 30 Years of Research and Practice. Oxford, UK/: Oxford University Press, 29–45.
Cohen, A. D. (2011a). Strategies in Learning and Using a Second Language: Research and Practice. Harlow, UK: Longman/ Pearson Education.
Cohen, A. D. (2011b). Test-taking strategies. In C. Coombe , P. Davidson , B. O'Sullivan and S. Stoynoff (eds.), The Cambridge Guide to Assessment. Cambridge, UK/: Cambridge University Press, 96–104.
Cohen, A. D. (2013). Verbal report. In C. A. Chapelle (ed.), The Encyclopedia of Applied Linguistics. Oxford, UK: Wiley-Blackwell. https://sites.google.com/a/umn.edu/andrewdcohen/publications/research-methodology (Retrieved from "Research Methodology" list, April 14, 2019).
Cohen, A. D. (2020). Considerations in assessing pragmatic appropriateness in spoken language. Language Teaching 53 183–202. https://doi.org/10.1017/S0261444819000156.
Cohen, A. D. and Upton, T. A. (2006). Strategies in Responding to the New TOEFL Reading Tasks (Monograph No. 33). Princeton, NJ: ETS.
https://sites.google.com/a/umn.edu/andrewdcohen/publications/assessment (Retrieved from "Assessment" list, April 14, 2019).
Cohen, A. D. and Wang, I. K.-H. (2018). Fluctuation in the functions of language learner strategies. System 74 169–182. https://doi.org/10.1016/j.system.2018.03.011.
Davidson, F. (2000). The language tester's statistical toolbox. System 28 605–617. https://doi.org/10.1016/S0346-251X(00)00041-5.
Dollerup, C. , Glahn, E. and Rosenberg Hansen, C. (1982). Reading strategies and test-solving techniques in an EFL-reading comprehension test – a preliminary report. Journal of Applied Language Study 1: 93–99.
Educational Testing Service (ETS) . (1995). TOEFL Practice Test B in TOEFL Practice Tests, vol. 1. Princeton, NJ: ETS, 55–106.
Educational Testing Service (ETS) . (2002). LanguEdge Courseware Score Interpretation Guide. Princeton, NJ: Educational Testing Service.
Fernandez, C. J. (2018). Behind a spoken performance: Test takers' strategic reactions in a simulated part 3 of the IELTS speaking test. Language Testing in Asia 8. Retrieved April 9, 2019, from https://doi.org/10.1186/s40468-018-0073-4.
Gavin, C. A. (1988). The Strategies of Native and Limited English Proficient Test-Takers as Revealed by Think Aloud Protocols. Unpublished EdD thesis, Rutgers University, New Brunswick, NJ.
Gordon, C. (1987). The Effect of Testing Method on Achievement in Reading Comprehension Tests in English as a Foreign Language. Unpublished master of arts thesis, Tel-Aviv University, Ramat-Aviv, Israel.
Guo, Q. , Kim, Y.-S. G. , Yang, L. and Liu, L. (2016). Does previewing answer choice options improve performance on reading tests? Reading and Writing 29 745–760. https://doi.org/10.1007/s11145-016-9626-z.
Homburg, T. J. and Spaan, M. C. (1981). ESL reading proficiency assessment: Testing strategies. In M. Hines and W. Rutherford (eds.), On TESOL'81. Washington, DC/: TESOL, 25–33.
Huang, H.-T. D. (2016). Exploring strategy use in L2 speaking assessment. System 63 13–27. https://doi.org/10.1016/j.system.2016.08.009.
Kashkouli, Z. , Barati, H. and Nejad Ansari, D. (2015). An investigation into the test-taking strategies employed for a high-stake test: Implications for test validation. International Journal of Research Studies in

Language Learning 4 61–72. https://doi.org/10.5861/ijrsll.2014.852.
Katalayi, G. B. (2018). Elimination of distracters: A construct-irrelevant strategy? An investigation of examinees' response decision processes in an EFL multiple-choice reading test. Theory and Practice in Language Studies 8 749–758. http://dx.doi.org/10.17507/tpls.0807.05.
Katalayi, G. B. and Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. Theory and Practice in Language Studies 3 877–884. https://doi.org/:10.4304/tpls.3.6.877-884.
Lumley, T. and Brown, A. (2004a). Test-taker and rater perspectives on integrated reading and writing tasks in the Next Generation TOEFL. Language Testing Update 35: 75–79.
Lumley, T. and Brown, A. (2004b). Test Taker Response to Integrated Reading/Writing Tasks in TOEFL: Evidence from Writers, Texts, and Raters. Final Report to ETS. Language Testing Research Centre, The University of Melbourne, Melbourne, Australia.
Nam, Y.-K. (2015). Korean EFL Students' Strategy Use in Gap-Filling Inference Items. English Teaching 70 81–107. https://doi.org/10.15858/engtea.70.4.201512.81.
Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. Language Testing 6 199–215. https://doi.org/10.1177/026553228900600206.
Nikolov, M. (2006). Test-taking strategies of 12-and 13-year-old Hungarian learners of EFL: Why whales have migraines. Language Learning 56 1–51. https://doi.org/10.1111/j.0023-8333.2006.00341.x.
Nyhus, S. E. (1994). Attitudes of Non-Native Speakers of English Toward the Use of Verbal Report to Elicit Their Reading Comprehension Strategies. Plan B Masters Paper, Department of ESL, University of Minnesota, Minneapolis.
O'Sullivan, B. and Weir, C. J. (2010). Language testing = validation. In B. O'Sullivan (ed.), Language Testing: Theories and Practices. Basingstoke, UK: Palgrave Macmillan.
O'Sullivan, B. and Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (ed.), Language Testing: Theories and Practice. Basingstoke, UK/: Palgrave Macmillan, 13–32.
Oxford, R. (2017). Teaching and Researching Language Learning Strategies: Self-Regulation in Context, 2nd edn. Abingdon, UK: Routledge.
Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. Modern Language Journal 91 323–340. https://doi.org/10.1111/j.1540-4781.2007.00583.x.
Poehner, M. E. (2008). Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting L2 Development. New York, NY: Springer.
Rogers, W. T. and Bateson, D. J. (1991). The influence of test-wiseness on the performance of high school seniors on school leaving examinations. Applied Measurement in Education 4 159–183. https://doi.org/10.1207/s15324818ame0402_5.
Russell, V. and Vásquez, C. (2018). Assessing the effectiveness of a web-based tutorial for interlanguage pragmatic development prior to studying abroad. IALLT Journal of Language Learning Technologies 48 69–96. Retrieved April 14, 2019, from https://journals.ku.edu/IALLT/article/view/8579.
Tian, S. (2000). TOEFL Reading Comprehension: Strategies Used by Taiwanese Students with Coachingschool Training. Unpublished PhD dissertation, Teachers College, Columbia University, New York.
Vahdany, F. , Akbari, E. , Shahrestani, F. and Askari, A. (2016). The relationship between cognitive and metacognitive strategy use and EFL listing test performance. Theory and Practice in Language Studies 6 385–391. http://dx.doi.org/10.17507/tpls.0602.22.
Wei, X. (2014). The intensity and direction of CET washback on Chinese college students' test- taking strategy use. Theory and Practice in Language Studies 4 1171–1177. https://doi.org/:10.4304/tpls.4.6.1171-1177.
Wu, A. D. and Stone, J. E. (2016). Validation through understanding test-taking strategies: An illustration with the CELPIP-General reading pilot test using structural equation modeling. Journal of Psychoeducational Assessment 34 362–379. https://doi.org/10.1177/0734282915608575.
Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. Language Testing 20 267–293. https://doi.org/10.1191/0265532203lt257oa.
Yang, P. (2000). Effects of Test-Wiseness Upon Performance on the Test of English as a Foreign Language. Unpublished PhD dissertation, University of Alberta, Edmonton, CN.
Youn, S. J. and Bi, N. Z. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. Intercultural Pragmatics 16 185–215. https://doi.org/10.1515/ip-2019-0009.

Prototyping new item types

Chapelle, C. , Enright, M. and Jamieson, J. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language. New York, NY: Routledge. This volume provides a detailed description of the research and development efforts to revise the TOEFL from 1990 to 2005. By describing this evolution, key principles in educational measurement are explained. The volume integrates the results of empirical studies and validity arguments that support the TOEFL iBT. Early chapters of the volume present the rationales for the revisions and a description of the process used to define the construct. Middle chapters provide detailed accounts of numerous research studies and prototyping efforts that informed the design of the test. The volume concludes with a validity argument for the test.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment, An Advanced Resource Book. New York, NY: Routledge. In the first section of this valuable resource book, the authors review issues of validity, test constructs, models, and frameworks. They consider the relationship between these abstract models, the development of a test framework, and actual test specifications. They explain how items and tasks are written and prototyped. They also consider ethics and standards for testing, test validation, and use. The second section of the book includes excerpts from highly influential papers by experts in these same areas. The third section focuses on group activities related to the core concepts of the book, such as analyzing items and tasks, creating arguments for test validation, and writing an assessment statement for a test.

Wolf, M. K. and Butler, Y. G. (eds.). (2017). English Language Proficiency Assessments for Young Learners. New York, NY: Routledge. This volume draws on examples of English language proficiency assessments for school-age children from the US and around the globe to highlight test development and validation processes. For those interested in examples of prototyping efforts, Chapters 3, 5, 7, and 10 may be of particulate note.

Bejar, I. , Douglas, D. , Jamieson, J. , Nissan, S. and Turner, J. (2000). TOEFL® 2000 Listening Framework: A Working Paper (TOEFL® Monograph No. MS-19). Princeton, NJ: Educational Testing Service.

Bowles, M. A. (2010). The Think-Aloud Controversy in Second Language Research. London, UK: Routledge.

Butler, F. , Eignor, D. , Jones, S. , McNamara, T. and Suomi, B. (2000). TOEFL® 2000 Speaking Framework: A Working Paper (TOEFL® Monograph No. MS-20). Princeton, NJ: Educational Testing Service.

Chapelle, C. , Enright, M. and Jamieson, J. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language. New York, NY: Routledge.

Cho, Y. and Choi, I. (2018). Writing from sources: Does audience matter? Assessing Writing 37 25–38. https://doi.org/10.1016/j.asw.2018.03.004.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. Language Assessment Quarterly 10 1–8. https://doi.org/10.1080/15434303.2011.622016.

Cumming, A. , Kantor, R. , Powers, D. , Santos, T. and Taylor, C. (2000). TOEFL® 2000 Writing Framework: A Working Paper (TOEFL® Monograph No. MS-18). Princeton, NJ: Educational Testing Service.

Enright, M. , Grabe, W. , Koda, K. , Mosenthal, P. , Mulcahy-Ernt, P. and Schedl, M. (2000). TOEFL® 2000 Reading Framework: A Working Paper. TOEFL Monograph Series MS – 17. RM-00–4. Princeton, NJ: Educational Testing Service.

Enright, M. and Schedl, M. (1999). Reading for a Reason: Using Reader Purpose to Guide Test Design (Internal Report). Princeton, NJ: Educational Testing Service.

Fulcher, G. (2003). Interface design in computer-based language testing. Language Testing 20 384–408. https://doi.org/10.1191/0265532203lt265oa.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment, An Advanced Resource Book. New York, NY: Routledge.

Harsch, C. and Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. Assessing Writing 17 228–250. https://doi.org/10.1016/j.asw.2012.06.003.

Jamieson, J. and Denner, A. (2000). Report on Prototype Project #7: Creation of Tasks and Observations for Listening Claim 2 – Listening for Pragmatic Understanding in a Variety of Text Types (Internal report). Princeton, NJ: Educational Testing Service.

Park, E. and Bredlau, E. (2014). Expanding the Question Formats of the TOEIC Speaking Test. (ETS White Paper). Princeton, NJ: Educational Testing Service.

Powers, D. E. , Roever, C. , Huff, K. L. and Trapani, C. S. (2003). Validating LanguEdge™ Courseware Scores Against Faculty Ratings and Student Self-assessments. (ETS Research Report Series): i–25.

Rosenfeld, M. , Leung, S. and Oltman, P. K. (2001). The Reading, Writing, Speaking, and Listening Tasks Important for Academic Success at the Undergraduate and Graduate Levels. TOEFL Monograph Series MS – 21. RM-01–03. Princeton, NJ: Educational Testing Service.

Pre-operational testing

Bachman, L. and Palmer, A. (2010). Language Assessment in Practice. Oxford, UK: Oxford University Press. This work describes the assessment use argument at length and helps contextualize the importance of pre-operational testing for various elements of a validation argument at large.

Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (2008). Building a validity argument for the Test of English as a Foreign LanguageTM. New York, NY: Routledge. This edited volume details the pre-operational testing for the revision of a large-scale, high-stakes test and discusses the implications for various inferences in the validation argument.

Fulcher, G. (2010). Practical Language Testing. New York, NY: Routledge. The author offers an entire chapter on the phases of pre-operational testing and provides plenty of examples and practical advice in expanding Fulcher and Davidson's (2007) initial description of pre-operational testing.

Lin, C. and MacGregor, D. (2018). Using a validation framework as a guide for planning analyses and collecting information in preoperational and operational testing. In J. Davis, J. Norris, M. Malone, T. McKay and Y. Son (eds.), Useful Assessment and Evaluation in Language Education. Washington, DC: Georgetown University Press, 201–216. This paper is based on the previous version of this handbook chapter and exemplifies how pre- operational testing issues can be situated in a validation argument.

AERA [American Educational Research Association], APA [American Psychological Association], NCME [National Council on Measurement in Education], Joint Committee on Standards for Educational and Psychological Testing (U.S.) . (2014). Standards for Educational and Psychological Testing. Washington, DC: AERA.

Alderson, J. C. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge, UK: Cambridge University Press.

ALTE [Association of Language Testers in Europe] . (2011). Manual for Language Test Development and Examining for Use with the CEFR. Strasbourg: Language Policy Division. Retrieved from www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford, UK: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (2010). Language Assessment in Practice. Oxford, UK: Oxford University Press.

Berry, V. , Nakatsuhara, F. , Inoue, C. and Galaczi, E. (2018). Exploring the use of video-conferencing technology to deliver the IELTS Speaking Test: Phase 3 technical trial. IELTS Partnership Research Papers, 2018/1. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-and-research/research-reports.

Bowles, M. (2010). The Think-Aloud Controversy in Second Language Research. New York, NY: Routledge.

Brown, J. D. (2001). Using Surveys in Language Programs. Cambridge, UK: Cambridge University Press.

Chapelle, C. , Enright, M. and Jamieson, J. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language™. New York, NY: Routledge. https://doi.org/10.4324/9780203937891.

Coombe, C. (2018). An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts. London, UK: British Council.

Corrigan, M. and Crump, P. (2015). Item analysis. Cambridge English: Research Notes 59: 4–9.

Davis, J. M. , Norris, J. M. , Malone, M. E. , McKay, T. H. and Son, Y. (2018). Useful Assessment and Evaluation in Language Education. Washington, DC: Georgetown University Press.

Davis, L. , Timpe-Laughlin, V. , Gu, L. and Ockey, G. (2018). Face-to-face speaking assessment in the digital age: Interactive speaking tasks online. In J. Davis , J. Norris , M. Malone , T. McKay and Y. Son (eds.), Useful Assessment and Evaluation in Language Education. Washington, DC/: Georgetown University Press, 115–130.

Dörnyei, Z. and Taguchi, T. (2009). Questionnaires in Second Language Research: Construction, Administration, and Processing (2nd ed.). New York, NY: Routledge. https://doi.org/10.4324/9780203864739

EALTA [European Association for Language Testing and Assessment] . (2006). Guidelines for Good Practice in Language Testing and Assessment. Retrieved from www.ealta.eu.org/documents/archive/guidelines/English.pdf.

Enright, M. K. , Bridgeman, B. , Eignor, D. , Kantor, R. N. , Mollaun, P. , Nissan, S. , Powers, D. E. and Schedl, M. (2008a). Prototyping new assessment tasks. In C. A. Chapelle , M. K. Enright and J. M. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language ™. New York, NY: Routledge, 97–144.

Enright, M. K. , Bridgeman, B. , Eignor, D. , Lee, Y.-W. and Powers, D. E. (2008b). Prototyping measures of listening, reading, speaking, and writing. In C. A. Chapelle , M. K. Enright and J. M. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language ™. New York, NY: Routledge, 145–186.

Fulcher, G. (2010). Practical Language Testing. New York, NY: Routledge.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. New York, NY: Routledge.

Gass, S. and Mackey, A. (2000). Stimulated Recall Methodology in Second Language Research. Mahwah, NJ: Lawrence Erlbaum Associates.
Gass, S. and Mackey, A. (2005). Second Language Research: Methodology and Design. Mahwah, NJ: Lawrence Erlbaum Associates.
Gass, S. and Mackey, A. (2007). Data Elicitation for Second and Foreign Language Research. New York: Routledge.
Green, R. (2013). Statistical Analyses for Language Testers. New York, NY: Palgrave Macmillan.
Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. Medical Care 38: 60–65.
Huff, K. , Powers, D. E. , Kantor, R. N. , Mollaun, P. , Nissan, S. and Schedl, M. (2008). Prototyping a new test. In C. A. Chapelle , M. K. Enright and J. M. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language ™. New York, NY: Routledge, 187–226.
ILTA [International Language Testing Association] . (2007). The ILTA Guidelines for Practice. Retrieved from www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf.
Kaplan, R. M. and Saccuzo, D. P. (1997). Psychological Testing: Principles, Applications and Issues. Pacific Grove, CA: Brooks Cole Publishing.
Kenyon, D. M. and MacGregor, D. (2012). Pre-operational testing. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. Abingdon, UK: Routledge.
Knoch, U. and Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. Language Testing 4 1–23. http://doi.org/10.1177/0265532217710049.
Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. London, UK: Longman.
Lin, C. and MacGregor, D. (2018). Using a validation framework as a guide for planning analyses and collecting information in preoperational and operational testing. In J. Davis , J. Norris , M. Malone , T. McKay and Y. Son (eds.), Useful Assessment and Evaluation in Language Education. Washington, DC/: Georgetown University Press, 201–216.
Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. The International Journal of Educational and Psychological Measurement 1 1–11. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1426043.
Nakatsuhara, F. , Inoue, C. , Berry, V. and Galaczi, E. (2016). Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery. A preliminary comparison of test-taker and examiner behaviour. IELTS Partnership Research Papers, 1. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-and- research/research-reports.
Nakatsuhara, F. , Inoue, C. , Berry, V. and Galaczi, E. (2017). Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and video-conferencing delivery (Phase 2). IELTS Partnership Research Papers, 3. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-and-research/research-reports.
Sireci, S. G. and Zenisky, A. L. (2016). Computerized innovative item formats. In S. Lane , M. R. Raymond , and T. M. Haladyna (eds.), Handbook of Test Development. New York, NY/: Routledge, 315–334.
Toulmin, S. E. (1958). The Uses of Argument. Cambridge: Cambridge University Press.
Wang, L. , Eignor, D. and Enright, M. K. (2008). A final analysis. In C. A. Chapelle , M. K. Enright and J. M. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign Language ™. New York, NY: Routledge, 259–318.
Wendler, C. and Walker, M. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. Downing and T. Haladyna (eds.), Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.
Wise, L. L. and Plake, B. S. (2016). Test design and development following the Standards for educational and psychological testing. In S. Lane , M. R. Raymond and T. M. Haladyna (eds.), Handbook of Test Development. New York, NY/: Routledge, 19–39.

Piloting vocabulary tests

Read, J. (2000). Assessing Vocabulary. Cambridge: Cambridge University Press. This volume in the authoritative Cambridge Language Assessment series was the first book-length treatment of second language vocabulary testing. It reviews relevant theory and research in language testing and second language acquisition, discusses the validity of four influential vocabulary tests, and offers practical advice on the issues involved in designing and scoring vocabulary measures for a variety of purposes.

Schmitt, N. (2010). Researching Vocabulary: A Vocabulary Research Manual. Basingstoke: Palgrave Macmillan. This book offers a comprehensive guide to research on vocabulary learning. It discusses both the substantive research questions to be investigated and the methodological issues that arise in designing good studies. There is a substantial section on measuring vocabulary knowledge which includes a critical review of many published vocabulary tests and guidance on how to develop new tests for various purposes. The author also outlines ten vocabulary research projects that new researchers could usefully undertake.

Webb, S. (ed.). (2020). The Routledge Handbook of Vocabulary Studies, London: Routledge. This is the most up-to-date and wide-ranging survey of research and professional practice on second language vocabulary. It includes a section of six chapters on measuring different aspects of vocabulary knowledge, along with another chapter by the present author which reviews key issues in designing vocabulary tests.

Zhang, D. and Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: Issues, findings and suggestions. Asia-Pacific Journal of Second and Foreign Language Education 2: 1–30. Zhang and Koda give a more complete account of all the research on the word associates format than has been possible in this chapter. It includes a systematic discussion of the design issues that have been investigated, as well as the procedures for administering the test, the different scoring methods and relevant characteristics of the test takers. The authors also suggest future directions for research on the format.

Alderson, J. C. , Haapakangas, E.-L. , Huhta, A. , Nieminen, L. and Ullakonoja, R. (2014). The Diagnosis of Reading in a Second or Foreign Language. New York, NY: Routledge.

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press.

Brown, H. D. and Abeywickrama, P. (2010). Language Assessment: Principles and Classroom Practices, 2nd edn. New York, NY: Pearson Longman.

Cremer, M. and Schoonen, R. (2013). The role of accessibility of semantic word knowledge in monolingual and bilingual fifth-grade reading. Applied Psycholinguistics 34 1195–1217. https://doi.org/10.1017/S0142716412000203.

Davies, A. , Brown, A. , Elder, C. , Hill, K. , Lumley, T. and McNamara, T. (1999). A Dictionary of Language Testing. Cambridge: Cambridge University Press.

Deese, J. (1965). The Structure of Associations in Language and Thought. Baltimore, MD: Johns Hopkins Press.

Fulcher, G. (2010). Practical Language Testing. London: Hodder Arnold.

Greidanus, T. , Beks, B. and Wakely, R. (2005). Testing the development of French word knowledge by advanced Dutch and English-speaking learners and native speakers. The Modern Language Journal 89 221–233. https://doi.org/10.3138/cmlr.62.4.509.

Greidanus, T. , Bogaards, P. , Van der Linden, E. , Nienhuis, L. and Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards and B. Laufer (eds.), Vocabulary in a Second Language: Selection, Acquisition, and Testing. Amsterdam: John Benjamins.

Greidanus, T. and Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. The Modern Language Journal 85 567–577. https://doi.org/10.1111/0026-7902.00126.

Gyllstad, H. (2020). Measuring knowledge of multiword items. In S. Webb (ed.), The Routledge Handbook of Vocabulary Studies. London: Routledge.

Harrington, M. (2018). Lexical Facility: Size, Recognition Speed and Consistency as Dimensions of Second Language Vocabulary Knowledge. London: Palgrave Macmillan.

Heaton, J. B. (1988). Writing English Language Tests, 2nd edn. London: Longman.

Hughes, A. (2003). Testing for Language Teachers, 2nd edn. Cambridge: Cambridge University Press.

Jang, E. E. (2014). Focus on Assessment. Oxford: Oxford University Press.

Kenyon, D. M. and MacGregor, D. (2012). Pre-operational testing. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London: Routledge.

Lado, R. (1961). Language Testing. London: Longman.

Meara, P. (2009). Connected Words: Word Associations and Second Language Vocabulary Acquisition. Amsterdam: John Benjamins.

Moussavi, S. A. (1999). A Dictionary of Language Testing, 2nd edn. Tehran: Rahnama Publications.

Nation, I. S. P. and Beglar, D. (2007). A vocabulary size test. The Language Teacher 31: 9–13.

Palermo, D. and Jenkins, J. (1964). Word Association Norms. Minneapolis, MN: University of Minnesota Press.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. Canadian Modern Language Review 56 282–308. https://doi.org/10.3138/cmlr.56.2.282.

Qian, D. D. and Schedl, M. (2004). Evaluating an in-depth vocabulary knowledge measure for assessing reading performance. Language Testing 21 28–52. https://doi.org/10.1191/0265532204lt273oa.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. Language Testing 10 355–371. https://doi.org/10.1177/026553229301000308.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (ed.), Validation in Language Assessment. Mahwah, NJ: Lawrence Erlbaum.
Read, J. (2000). Assessing Vocabulary. Cambridge: Cambridge University Press.
Reed, D. J. (2014). Field testing of test items and tasks. In A. J. Kunnan (ed.), The Companion to Language Assessment. Malden, MA: Wiley Blackwell.
Schmitt, N. , Ng, J. W. C. and Garras, J. (2011). The word associates format: Validation evidence. Language Testing 28 105–126. https://doi.org/10.1177/0265532210373605.
Schoonen, R. and Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. Language Testing 25 211–236. https://doi.org/10.1177/0265532207086782.
Spolsky, B. (1995). Measured Words. Oxford: Oxford University Press.
Verhallen, M. and Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. Applied Linguistics 14 344–365. https://doi.org/10.1093/applin/14.4.344.
Wesche, M. and Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Breadth vs. depth. Canadian Modern Language Review 53: 13–39.
Xue, G. and Nation, I. S. P. (1984). A university word list. Language Learning and Communication 3: 215–229.
Zhang, D. , Yang, X. , Lin, C.-H. and Gu, Z. (2017). Developing a word associates test to assess L2 Chinese learners' vocabulary depth. In D. Zhang and C.- H. Lin (eds.), Chinese as a Second Language Assessment. Singapore: Springer.

Classical test theory

Brown, J. D. (2005). Testing in Language Programs: A Comprehensive Guide to English Language Assessment, new edn. New York, NY: McGraw-Hill. This book provides an introduction to language testing for language teachers and teachers in training. To do so, it covers both norm-referenced CTT and criterion-referenced testing and provides straightforward explanations of all concepts and clear instructions for calculating all the related statistics by hand or using an Excel spreadsheet (with written instructions only).
Carr, N. T. (2011). Designing and Analyzing Language Tests. Oxford: Oxford University Press. This book also presents a general introduction to language testing. It focuses primarily on CTT and provides explanations for how to calculate all the related statistics by hand but also provides effective video modules that walk readers through the process of using an Excel spreadsheet to calculate the related statistics.
Hauenstein, C. E. and Embretson, S. E. (2018). Classical test theory. In B. B. Frey (ed.), The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. Thousand Oaks, CA: Sage. This encyclopedia entry provides a short and accessible introduction to key theoretical issues in CTT, including true-score and error variances, the four most commonly used reliability estimates, the standard error of estimate, item analysis, validity, and the limitations of CTT.
Traub, R. E. (1997). Classical test theory in historical perspective. Educational Measurement Issues and Practice 16: 8–14. This article offers a relatively short and clear historical overview of the development of CTT beginning with background from the beginning of the twentieth century and then coverage of the key achievements in CTT over the years, discussion of how CTT was formalized with references from 1923 to 1968, and then concluding remarks focusing on the problems and limitations of CTT.
Alderson, C. J. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge, UK: Cambridge University.
Association of Language Testers in Europe . (1998). Studies in Second Language Testing 6: Multilingual Glossary of Language Testing Terms. Cambridge, UK: Cambridge University Press.
Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford, UK: Oxford University Press.
Brown, J. D. (2005). Testing in Language Programs: A Comprehensive Guide to English Language Assessment, new edn. New York, NY: McGraw-Hill.
Brown, J. D. and Bailey, K. M. (2008). Language testing courses: What are they in 2007. Language Testing 25 349–383. https://doi.org/10.1177/0265532208090157.
Brown, J. D. and Hudson, T. (2002). Criterion-Referenced Language Testing. Cambridge, UK: Cambridge University Press.
Brown, J. D. and Ross, J. A. (1996). Decision dependability of item types, sections, tests and the overall TOEFL test battery. In M. Milanovic and N. Saville (eds.), Performance Testing, Cognition and Assessment. Cambridge, UK: Cambridge University Press.
Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology 3 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

DeVellis, R. F. (2006). Classical test theory. Medical Care 44: S50–S59.
Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika 10: 255–282.
Henning, G. (1984). Advantages of latent trait measurement in language testing. Language Testing 1 123–133. https://doi.org/10.1177/026553228400100201.
Hinofotis, F. B. (1981). Perspectives on language testing: Past, present and future. Nagoya Gakuin Daigaku Gaikokugo Kyoiku Kiyo 4: 51–59.
Holland, P. W. and Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. Psychometrika 68: 123–149.
Hughes, A. (2002). Testing for Language Teachers. Cambridge, UK: Cambridge University Press.
Kuder, G. F. and Richardson, M. W. (1937). The theory of estimation of test reliability. Psychometrika 2 152–260. https://doi.org/10.1007/BF02288391.
Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. New York, NY: McGraw-Hill.
Lord, F. M. and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher 18 5–11. https://doi.org/10.3102/0013189X018002005.
Messick, S. (1996). Validity and washback in language testing. Language Testing 13 241–256. https://doi.org/10.1177/026553229601300302.
Pearson, K. (1896). Mathematical contributions to the theory of evolution-III. Regression, heredity and panmixia. Philosophical Transactions, A 187 252–318. https://doi.org/10.1098/rsta.1896.0007.
Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology 15 72–101. https://doi.org/10.1037/11491-005.
Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology 3 171–195. http://proquest.umi.com/login/athens?url=www.proquest.com/scholarly-journals/correlation-calculated-faulty-data/docview/1293688112/se-2?accountid=11979.
Spolsky, B. (1978). Introduction: Linguists and language testers. In B. Spolsky (ed.), Advances in Language Testing Series: 2. Arlington, VA: Center for Applied Linguistics.

Item response theory and many-facet Rasch measurement

Bond, T. C. and Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd edn. Mahwah, NJ: L. Erlbaum. The authors provide a transparent introduction to the Rasch model. The book is written for an audience without advanced mathematical knowledge. Various examples are described, and many are accompanied by interpreted computer output. An emphasis of the book is to convince readers that the 1PL Rasch model is appropriate for all situations.
Eckes, T. (2015). Introduction to Many-Facet Rasch Measurement. New York, NY: Peter Lang. In this book, Eckes masterfully demonstrates appropriate uses of MFRM and the FACETS software program for language assessment practitioners and researchers. Eckes makes challenging information transparent for a non-technical language assessment audience. This book can be used as a textbook for language assessment graduate courses.
Green, R. (2013). Statistical Analyses for Language Testers. Basingstoke: Palgrave Macmillan. In this book, Green begins with an overview of classical test theory (CTT) statistical analyses procedures before providing a light introduction to item response theory (IRT). Green does a nice job of comparing principles of IRT with those of CTT. This book is appropriate for readers who have little or no background knowledge in statistical approaches to language assessment analyses and want a basic knowledge of both CTT and IRT for language assessment.
McNamara, T. , Knoch, U. and Fan, J. (2019). Fairness, Justice, and Language Assessment. Oxford: Oxford University Press. This book, which is a revised version of McNamara's (1996) *Measuring Second Language Performance*, is a must-read for language assessment researchers who are interested in performance-based assessments and/or MFRM. The book is both conceptual and practical. It provides a very clear introduction to the principles underlying the Rasch model as well as interpretations and explanations of output from the FACETS program. It demonstrates how to apply MFRM and the FACETS software program on contemporary issues of fairness in language assessment.
Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. Language Assessment Quarterly 13 1–24. https://doi.org/10.1080/15434303.2015.1133626.
Batty, A. (2018). Investigating the impact of nonverbal communication cues on listening item types. In G. J. Ockey and E. Wagner (eds.), Assessing L2 Listening: Moving Towards Authenticity. Philadelphia, PA/: John Benjamins, 161–175.

Bock, R. D. (1997). A brief history of item response theory. Educational Measurement: Issues and Practice 16 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x.
Bond, T. C. and Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd edn. Mahwah, NJ: L. Erlbaum.
Bonk, W. and Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. Language Testing 20: 1, 89–110. https://doi.org/10.1191/0265532203lt245oa.
Davidson, F. and Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing 2 164–179. https://doi.org/10.1177/026553228500200205.
de Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. New York, NY: The Guilford Press.
Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. Language Assessment Quarterly 2 197–221. https://doi.org/10.1207/s15434311laq0203_2.
Eckes, T. (2015). Introduction to Many-Facet Rasch Measurement. New York, NY: Peter Lang.
Elder, C. , Knoch, U. , Barkhuizen, G. and von Radow, J. (2005). Individual feedback to enhance rater training: Does it work? Language Assessment Quarterly 2 175–196. https://doi.org/10.1207/s15434311laq0203_1.
Ferne, T. and Rupp, A. (2007). Synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. Language Assessment Quarterly 4 113–148. https://doi.org/10.1080/15434300701375923.
Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. Language Testing 13 23–51. https://doi.org/10.1177/026553229601300103.
Garret, N. (1991). Technology in the services of language learning: Trends and issues. The Modern Language Journal 75 74–101. https://doi.org/10.1111/j.15404781.2009.00968.x.
Green, R. (2013). Statistical Analyses for Language Testers. Basingstoke: Palgrave Macmillan.
He, T. H. , Gou, W. J. , Chien, Y. C. , Chen, I. , Shan, J. and Chang, S. M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. Psychological Reports 112 469–486. https://doi.org/10.2466/03.11.PR0.112.2.469-485.
Henning, G. (1984). Advantages of latent trait measurement in language testing. Language Testing 1 123–133. https://doi.org/10.1177/026553228400100201.
Henning, G. , Hudson, T. and Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. Language Testing 2 41–154. https://doi.org/10.1177/026553228500200203.
Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. Language Testing 28 179–200. https://doi.org/10.1177/0265532210384252.
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing 19 3–31. https://doi.org/10.1191/0265532202lt218oa.
Koyama, D. , Sun, A. and Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. Language Learning & Technology 20 148–165. http://llt.msu.edu/issues/february2016/koyamasunockey.pdf.
Li, S. (2019). Variations in rating scale functioning in assessing speech act production in L2 Chinese. Language Assessment Quarterly 16 271–293. https://doi.org/10.1080/15434303.2019.1648473.
Linacre, J. M. (1987–2019). Many Facets Rasch Measurement, Version 3.83.0. Chicago, IL: Mesa Press.
Lord, F. M. (1952). A Theory of Test Scores. Psychometric Monograph No. 7. Richmond, VA: Psychometric Corporation.
Lord, F. M. and Novick, M. (1968). Statistical Theories of Mental Test Scores, Reading, MA: Addison-Wesley.
McNamara, T. (1996). Measuring Second Language Performance. New York, NY: Addison Wesley Longman Limited.
McNamara, T. and Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. Language Testing 29 555–576. https://doi.org/10.1177/0265532211430367.
McNamara, T. , Knoch, U. and Fan, J. (2019). Fairness, Justice, and Language Assessment. Oxford: Oxford University Press.
Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. Language Assessment Quarterly 4 149–164. https://doi.org/10.1080/15434300701375717.
Ockey, G. J. and Choi, I. (2015). Item response theory. In C. A. Chapelle (ed.), Encyclopedia of Applied Linguistics. Oxford: John Wiley and Sons, Inc.
Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago Press.
Thurston, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology 16 433–451. https://doi.org/10.1037/h0073357.

Weigle, S. (1998). Using FACETS to model rater training effects. Language Testing 15 263–287. https://doi.org/10.1177/026553229801500205.
Woods, A. and Baker, R. (1985). Item response theory. Language Testing 2 117–140. https://doi.org/10.1177/026553228500200202.

Reliability and dependability

Brown, J. D. (1989). Criterion-referenced test reliability. University of Hawai'i Working Papers in ESL 1: 79–113. This article explains in detail the definitions and operationalizations of reliability in both norm-referenced and criterion-referenced tests, with a focus on the latter. It is a good introductory source on test reliability.
Dimova, S. , Yan, X. and Ginther, A. (2020). Local Language Testing: Design, Implementation, and Development. New York, NY: Routledge. This book does not focus on reliability per se. However, it provides a non-technical account of assessment principles, including item and rater reliability, through assessment activities in local contexts. This book is useful for language teachers, program coordinators, and novice language testing researchers who are interested in obtaining a general understanding of reliability in local assessment practices.
Krzanowski, W. J. and Woods, A. J. (1984). Statistical aspects of reliability in language testing. Language Testing 1: 1–20. This article discusses the statistical properties of commonly used reliability estimates and explains how those estimates can be derived from ANOVA models.
McNamara, T. , Knoch, U. and Fan, J. (2019). Fairness, Justice and Language Assessment. Oxford: Oxford University Press. This book offers a practical account of how analyses of item, rater, and test reliability using Rasch measurement models can provide evidence for test fairness. Different from most articles or books on test reliability, this book integrates psychometric reliability in building fairness argument for language tests.
Myford, C. M. and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement 4: 386–422.
Myford, C. M. and Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. Journal of Applied Measurement 5: 189–227. These two articles provide a conceptual explanation of rater effects and illustrate how these rater effects can be examined through many-facet Rasch measurement models. It is a useful source as an introduction to Rasch models and rater reliability.
AERA, APA, & NCME . (2014). Standards for Educational and Psychological Testing. Washtington, DC: AERA.
Alderson, J. C. (2010). Cognitive diagnosis and q-matrices in language assessment: A commentary. Language Assessment Quarterly 7 96–103. https://doi.org/10.1080/15434300903426748.
Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford, UK: Oxford University Press.
Bennett, R. (2010, February 18). Formative Assessment: A Critical Review. Cambridge, UK: Cambridge Assessment Network Seminar.
Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement 38 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x.
Brown, J. D. (1989). Criterion-referenced test reliability. University of Hawai'i Working Papers in English as a Second Language 8: 79–113.
Brown, J. D. and Hudson, T. (2002). Criterion-Referenced Language Testing. Cambridge, UK: Cambridge University Press.
Chapelle, C. A. , Chung, Y. R. , Hegelheimer, V. , Pendar, N. and Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. Language Testing 27 443–469. https://doi.org/10.1177/0265532210367633.
Coe, R. J. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch model. Oxford Review of Education 34 609–636. https://doi.org/10.1080/03054980801970312.
Coleman, J. A. (2004). Modern languages in British universities: Past and present. Arts and Humanities in Higher Education 3 147–162. https://doi.org/10.1177/1474022204042684.
Council of Europe . (2001). Common European Framework of Reference for Languages, Teaching, Learning, Assessment. Cambridge, UK: Cambridge University Press.
Ennis, R. H. (1999). Test reliability: A practical exemplification of ordinary language philosophy. In R. Curren (ed.), Philosophy of Education. Urbana, IL/: The Philosophy of Education Society, 242–248.
Fan, J. and Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? Papers in Language Testing and Assessment 8: 117–142.
Feldt, L. S. and Brennan, R. L. (1989). Reliability. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Macmillan, 105–146.

Ferris, T. L. J. (2004). A new definition of measurement. Measurement 36 101–109. https://doi.org/10.1016/j.measurement.2004.03.001.
Frederiksen, N. , Mislevy, R. J. and Bejar, I. (eds.). (1993). Test Theory for a New Generation of Tests. Hillsdale, NJ: Erlbaum.
Gulliksen, H. (1950). Theory of Mental Tests. New York, NY: Wiley.
Jacobs, H. , Zinkgraf, S. , Wormuth, D. , Hartfiel, V. and Hughey, J. (1981). Testing ESL Composition: A Practical Approach. Rowley, MA: Newbury House.
Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin 12 527–535. https://doi.org/10.1037/0033-2909.112.3.527.
Knoch, U. and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. Language Testing 35 477–499. https://doi.org/10.1177/0265532217710049.
Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. Language Testing 28 543–560. https://doi.org/10.1177/0265532211406422.
McNamara, T. and Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. Language Testing 29 555–576. https://doi.org/10.1177/0265532211430367.
McNamara, T. , Knoch, U. and Fan, J. (2019). Fairness, Justice and Language Assessment. Oxford, UK: Oxford University Press.
McNamara, T. and Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. Language Assessment Quarterly 8 161–178. https://doi.org/10.1080/15434303.2011.565438.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: American Council on Education and Macmillan, 1–103.
Mislevy, R. J. (1994). Can There be Reliability Without ‘Reliability’? Princeton, NJ: Educational Testing Service.
Mislevy, R. J. (2004). Can there be reliability without ‘reliability’? Journal of Educational and Behavioral Statistics 29: 241–244.
Mislevy, R. J. , Almond, R. G. and Lukas, J. F. (2003). A Brief Introduction to Evidence-Centred Design. Research Report RR-03-16. Princeton, NJ: Educational Testing Service.
Moss, P. (1994). Can there be validity without reliability? Educational Researcher 23 5–12. https://doi.org/10.3102/0013189X023002005.
Moss, P. (2004). The meaning and consequences of ‘reliability.’ Journal of Educational and Behavioral Statistics 29: 245–249.
Shavelson, R. J. (2008). Guest editor’s introduction. Applied Measurement in Education 21 293–294. https://doi.org/10.1080/08957340802347613.
Spolsky, B. (1995). Measured Words: Development of Objective Language Testing. Oxford, UK: Oxford University Press.
Stellman, J. M. (1998). Encyclopedia of Occupational Health and Safety, 4th edn. Geneva, Switzerland: International Labor Office.
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approached to estimating integrated reliability. Practical Assessment, Research & Evaluation 9 1–11. https://doi.org/10.7275/96jp-xz07.
Weigle, S. C. (2002). Assessing Writing. Cambridge, UK: Cambridge University Press.
Winke, P. and Lim, H. (2015). ESL essay raters’ cognitive processes in applying the Jacobs *et al.* rubric: An eye-movement study. Assessing Writing 25 38–54. https://doi.org/10.1016/j.asw.2015.05.002.
Xi, X. (2010). Aspects of performance online graph description tasks: Influenced by graph familiarity and different task features. Language Testing 27 73–100. https://doi.org/10.1177/0265532209346454.
Yan, X. , Cheng, L. and Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. Language Testing 36 207–234. https://doi.org/10.1177/0265532218775764.
Yan, X. and Ginther, A. (2017). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang and A. Ginther (eds.), Assessment in L2 Pronunciation. Routledge, 67–88.

Scoring performance tests

Crusan, D. (2015). The use of rubrics to assess writing: Issues and challenges. Assessing Writing 26: 1–82. A special issue focusing on the development and revision of assessment scales for a range of contexts.

East, M. and Cushing, S. (2016). Innovation in rubric use: Exploring different dimensions. Assessing Writing 30: 1–76. A special issue offering discussions on different contexts of use of assessment scales.

Kuiken, F. and Vedder, I. (2014). Assessing oral and written L2 performance: Raters' decisions, rating procedures and rating scales. Language Testing 31: 285–392. A special issue exploring rater-focused questions.

Wind, S. A. and Engelhard, G. (2019). Rater-mediated assessments. Journal of Educational Measurement 56: 473–663. A special issue focusing on raters from fields beyond L2 assessment.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson and B. North (eds.), Language Testing in the 1990s. London/: Modern English Publications/British Council, 71–86.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education . (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater v. 2. Journal of Technology, Learning and Assessment 4 1–31. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. Language Assessment Quarterly 9 225–248. https://doi.org/10.1080/15434303.2011.637262.

Ballard, P. B. (1923). Mental Tests. London, England: Hodder and Stoughton.

Banerjee, J. , Yan, X. , Chapman, M. and Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. Assessing Writing 26 5–19. https://doi.org/10.1016/j.asw.2015.07.001.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. Assessment in Education: Principles, Policy & Practice 18 279–293. https://doi.org/10.1080/0969594X.2010.526585.

Barnwell, D. (1986). A History of Foreign Language Testing in the United States. Tempe, AZ: Bilingual Press.

Bejar, I. (2012). Rater cognition: Implications for validity. Educational Measurement: Issues and Practice 31 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. Journal of Applied Measurement 6: 202–223.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman and A. D. Cohen (eds.), Interfaces Between Second Language Acquisition and Language Testing Research. New York, NY/: Cambridge University Press, 112–140.

Broad, B. (2003). What We Really Value: Beyond Rubrics in Teaching and Assessing Writing. Logan, UT: Utah State University Press.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London/: Routledge, 413–425.

Brown, J. D. , Hudson, T. D. , Norris, J. M. and Bonk, W. (2002). An Investigation of Second Language Task-Based Performance Assessments. Honolulu, HI: University of Hawai'i Press.

Burstein, J. and Chodorow, M. (1999). Automated Essay Scoring for Nonnative English Speakers. Retrieved from www.aclweb.org/anthology/W99-0411.pdf.

Caliskan, A. , Bryson, J. J. and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science 356, 183–186. https://doi.org/10.1126/science.aal4230.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), Interfaces Between Second Language Acquisition and Language Testing Research. New York, NY/: Cambridge University Press, 32–70.

Clark, J. L. D. (1978). Interview testing research at educational testing service. In J. L. D. Clark (ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ/: Educational Testing Service, 211–229.

Congdon, P. J. and McQueen, J. (2005). The stability of rater severity in large-scale assessment programs. Journal of Educational Measurement 37 163–178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x.

Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Davies, A. , Brown, A. , Elder, C. , Hill, K. , Lumley, T. and McNamara, T. (1999). Dictionary of Language Testing. Studies in Language Testing volume 7. Cambridge: UCLES/Cambridge University Press.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. Journal of the Royal Statistical Society 53 644–663. Retrieved from www.jstor.org/stable/2979547.

Elder, C. , McNamara, T. , Woodward-Kron, R. , Manias, E. , McColl, G. , Webb, G. and Pill, J. (2013). Toward Improved Healthcare Communication: Development and Validation of Language Proficiency Standards for Non-Native English Speaking Health Professionals. Melbourne: University of Melbourne.

Erdosy, M. U. (2004). Exploring validity in judging writing ability in a second language: A study of four experienced raters of ESL compositions. In TOEFL Research Report No TOEFL-RR-70. Princeton, NJ: Educational Testing Service.
European Commission . (2012). First European Survey on Language Competences: Technical Report. Luxembourg: Publications Office of the European Union.
Fairbairn, J. and Dunlea, J. (2017). Speaking and Writing Rating Scales Revision Technical Report. London: British Council Assessment Research Group.
Fechner, G. T. (1897). Kollektivmasslehre. Leipzig, Germany: Wilhelm Engelmann.
Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. English Language Teaching Journal 41 287–291. https://doi.org/10.1093/elt/41.4.287.
Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. Language Testing 13 208–238. https://doi.org/10.1177/0265532209359514.
Fulcher, G. (2003). Testing Second Language Speaking. London: Longman Pearson Education.
Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy and N. H. Hornberger (eds.), Encyclopedia of Language and Education, Language Testing and Assessment, 2nd edn., vol. 7. New York, NY: Springer, 157–176.
Fulcher, G. (2009). Test use of political philosophy. Annual Review of Applied Linguistics 29 3–20. https://doi.org/10.0.3.249/S0267190509090023.
Fulcher, G. (2012). Scoring performance tests. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. Abingdon and New York, NY/: Routledge, 378–392.
Fulcher, G. , Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. Language Testing 28 5–29. https://doi.org/10.1177/0265532209359514.
Galaczi, E. D. and Khabbazbashi, N. (2016). Rating scale development: A multi-stage exploratory sequential design. In J. Creswell , A. Moeller and N. Saville (eds.), Second Language Assessment and Mixed Methods Research. Studies in Language Testing, vol. 43. Cambridge: UCLES/Cambridge University Press, 208–232.
Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (ed.), Assessing Second Language Writing in Academic Contexts. Norwood/: Ablex, 241–276.
Hamp-Lyons, L. (2016). Farewell to holistic scoring? Assessing Writing 27: A1–A2. https://doi.org/10.1016/j.asw.2015.12.002.
Harsch, C. and Seyferth, S. (2020). Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist. Assessing Writing 43 1–15. https://doi.org/10.1016/j.asw.2019.100433.
Hirai, A. and Koizumi, H. (2013). Validation of empirically derived rating scales for a story retelling speaking test. Language Assessment Quarterly 10 398–422. https://doi.org/10.1080/15434303.2013.824973.
Hulstijn, J. A. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. Modern Language Journal 91 663–667. https://doi.org/10.0.4.87/j.1540-4781.2007.00627_5.x.
Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson and B. A. Huot (eds.), Validating Holistic Scoring for Writing Assessment. Cresskill, NJ/: Hampton Press, 206–232.
Isaacs, R. , Trofimovich, P. , Yu, G. and Munoz Chereau, B. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. IELTS Research Reports, 2015–4.
Jacoby, S. (1998). Science as Performance: Socializing Scientific Discourse Through the Conference Talk Rehearsal. Unpublished doctoral dissertation. University of California, Los Angeles.
Jones, N. (2009). A comparative approach to constructing a multilingual proficiency framework: Constraining the role of standard-setting. In N. Figueras and J. Noijons (eds.), Linking to the CEFR Levels: Research Perspectives. Arnhem/: CITO/EALTA, 35–44.
Kane, M. (2006). Validation. In R. L. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: American Council on Education, Praeger Publishers, 17–64.
Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. The Modern Language Journal 28 136–150. https://doi.org/10.1111/j.1540-4781.1944.tb04835.x.
Khabbazbashi, N. and Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. Language Testing, 1–28. https://doi.org/10.1177/0265532219898635.
Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. Language Testing 28 543–560. https://doi.org/10.1177/0265532211406422.
Lim, G. S. (2012). Developing and validating a mark scheme for writing. Research Notes 49: 6–10.
Linacre, J. M. (1989). Many Facet Rasch Measurement. Chicago, IL: MESA Press.
Lumley, T. (2005). Assessing Second Language Writing: The Rater's Perspective. Frankfurt: Peter Lang.
Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. English Language Teaching Journal 44 117–121.

https://doi.org/10.1093/elt/44.2.117.
McNamara, T. F. (1996). Measuring Second Language Performance. London: Longman.
Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63 81–97. https://doi.org/10.1037/h0043158.
Mislevy, R. J. , Almond, R. G. and Lukas, J. F. (2003). A Brief Introduction to Evidence-Centered Design. Research Report RR-03–16. Princeton, NJ: Educational Testing Service.
Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher 23 5–12. https://doi.org/10.3102/0013189X023002005.
Myford, C. M. and Wolfe, A. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement 4: 386–422.
Norris, J. M. , Brown, J. D. , Hudson, T. and Bonk, J. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. Language Testing 19 395–418. https://doi.org/10.1191/0265532202lt237oa.
North, B. (2000). The Development of a Common Framework Scale of Language Proficiency. Frankfurt: Peter Lang.
Oscarson, M. (1980). Approaches to Self-Assessment in Foreign Language Learning. Oxford: Pergamon.
Oscarson, M. (1989). Self-Assessment of Language Proficiency: Rationale and Applications. Language Testing 6 1–13. https://doi.org/10.1177/026553228900600103.
Pollitt, A. (1991). Giving students a sporting chance: Assessment by counting and judging. In C. Alderson and B. North (eds.), Language Testing in the 1990s. London/: Macmillan, 46–59.
Roach, J. O. (1936). The reliability of school certificate results. Overseas Education: A Journal of Educational Experiment and Research in Tropical and Subtropical Areas 7: 113–118.
Roach, J. O. (1945). Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students. University of Cambridge Local Examinations Syndicate Report Circulated to Oral Examiners and Local Examiners.
Ross, S. (2012). Claims, evidence, and interference in performance assessment. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. Abingdon and New York, NY/: Routledge, 223–233.
and Şahan Ő Razi, S. (2020). Do experience and text quality matter for raters' decision-making behaviours? Language Testing, 1–22. https://doi.org/10.1177/0265532219900228.
Schmidt, E. and Pastorino, C. (forthcoming). Eye-tracking and EEG in language assessment. In G. Yu and J. Xu (eds.), Language Test Validation in a Digital Age. Studies in Language Testing Volume 52. Cambridge: UCLES/Cambridge University Press.
Spolsky, B. (1995). Measured Words. Oxford: Oxford University Press.
Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), Examining Speaking: Research and Practice in Assessing Second Language Speaking. Studies in Language Testing Volume 30. Cambridge/: UCLES/Cambridge University Press, 171–233.
Thorndike, E. L. (1912). Education: A First Book. New York, NY: Macmillan.
Upshur, J. and Turner, C. E. (1995). Constructing rating scales for second language tests. English Language Teaching Journal 49 3–12. https://doi.org/10.1093/elt/49.1.3.
Wang, X. , Evanini, K. , Zechner, K. and Mulholland, M. (2017). Modeling discourse coherence for the automated scoring of spontaneous spoken responses. Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education, 132–137.
Weigle, S. C. (2002). Assessing Writing. Cambridge: Cambridge University Press.
Weir, C. J. (2005). Limitations of the Common European framework for developing comparable examinations and tests. Language Testing 22 281–300. https://doi.org/10.1191/0265532205lt309oa.
Weir, C. J. , Vidaković, I. and Galaczi, E. D. (2013). Measured Constructs: A History of Cambridge English Language Examinations 1913–2012. Studies in Language Testing Volume 37. Cambridge: UCLES/Cambridge University Press.
Wesolowski, B. C. (2019). Predicting operational rater-type classifications using Rasch Measurement theory and Random Forests: A music performance assessment perspective. Journal of Educational Measurement 56 610–625. https://doi.org/10.1111/jedm.12227.
White, E. M. (1984). Holisticism. College Composition and Communication 35: 400–409.
Wilds, C. (1979). The measurement of speaking and reading proficiency in a foreign language. In M. L. Adams and J. R. Firth (eds.), Testing kit: French and Spanish. U.S. Department of State, Arlington, VI/: Foreign Services Institute, 1–12.
Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test (TAST) for operational use. Language Testing 24 251–286. https://doi.org/10.1177/0265532207076365.
Yannakoudakis, H. , Briscoe, T. and Alexopoulou, T. (2012). Automating Second Language Acquisition Research: Integrating Information Visualization and Machine Learning. Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH, 35–43. Retrieved from www.aclweb.org/anthology/W12-0206.pdf.

Yannakoudakis, H. , Briscoe, T. and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 180–189. Retrieved from www.aclweb.org/anthology/P11-1019.pdf.
Yapo, A. and Weiss, J. (2018). Ethical implications of bias in machine learning. In Proceedings of the 51st Hawaii International Conference on System Sciences. Retrieved from scholarspace.manoa.hawaii.edu/bitstream/10125/50557/paper0670.pdf.
Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. Assessing Writing 27 37–53. https://doi.org/10.1016/j.asw.2015.11.001.

Validity and the automated scoring of performance tests

Bennett, R. E. and Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice 17: 9–17. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x. This article provides the first comprehensive treatment of automated scoring as an integral part of an assessment, which is a departure from seeing it as merely a substitute for human scoring. A central argument in this article is that automated scoring should be designed as a dynamic component in the assessment process, interacting with the construct definition, test and task design, test taker interface, and reporting methods.
Clauser, B. E. , Kane, M. T. and Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. Applied Measurement in Education 15: 413–432. https://doi.org/10.1207/S15324818AME1504_05. This is the first attempt to integrate automated scoring into an argument-based validation framework, thus highlighting the complex and expanded effects that automated scoring may have on the overall validity argument for the entire assessment. It discusses potential validity threats to the strength of each inference in the validity argument that may be introduced by automated scoring, pointing to the critical research that is needed to discount or reduce the threats.
Xi, X. (2010). Automated scoring and feedback systems – Where are we and where are we heading? Language Testing 27: 291–300. https://doi.org/10.1177/0265532210364643. This introduction to a special issue offers a critical review of the research and development work in automated scoring and feedback systems for language assessment and learning. It raises a series of relevant validity questions for automated scoring and feedback systems, respectively. These validity questions are linked to the inferences in the argument-based validation framework. It also provides a brief overview of the seven articles featured in the special issue.
Zechner, K. and Evanini, K. (eds.). (2020). Automated speaking assessment: Using language technologies to score spontaneous speech. In J. Norris, S. Ross, S. Weigle and X. Xi (eds.), Innovations in Language Learning and Assessment at ETS. New York, NY: Routledge. This volume provides a state-of-the-art review of research and development of ETS's speech technologies situated in the broader context of developments and advances in the field. It attempts to open the black box of automated speech scoring by offering an under-the-hood analysis of the components of automated speech scoring systems, including the speech recognition system, speech features, and speech scoring models. It also discusses validity issues involved in automated speech scoring and provides an overview of recent developments and future outlook.
Attali, Y. , Bridgeman, B. and Trapani, C. (2010). Performance of a generic approach in automated essay scoring. The Journal of Technology, Learning, and Assessment 10. Retrieved December 21, 2020, from https://files.eric.ed.gov/fulltext/EJ895962.pdf.
Bennett, R. E. (1993). On the meaning of constructed response. In R. E. Bennett and W. C. Ward (eds.), Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment. Hillsdale, NJ/: Lawrence Erlbaum Associates, 1–27.
Bennett, R. E. and Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice 17 9–17. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x.
Bernstein, J. , Van Moere, A. and Cheng, J. (2010). Validating automated speaking tests. Language Testing 27 355–377. https://doi.org/10.1177/0265532210364404.
Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. Language Learning and Technology 5 95–97. Retrieved December 21, 2020, from http://llt.msu.edu/vol5num2/deville/default.html.
Chapelle, C. A. and Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment [Special issue]. Language Testing 27 301–315. https://doi.org/10.1177/0265532210364405.
Chapelle, C. A. , Cotos, E. and Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language Testing 32 385–405. https://doi.org/10.1177/0265532214565386.
Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign LanguageTM. Mahwah, NJ: Lawrence Erlbaum.

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. English Journal 93: 47–52.
Clauser, B. E. , Kane, M. T. and Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. Applied Measurement in Education 15 413–432. https://doi.org/10.1207/S15324818AME1504_05.
Clauser, B. E. , Swanson, D. B. and Clyman, S. G. (2000). A comparison of the generalizability of scores produced by expert raters and automated scoring systems. Applied Measurement in Education 12 281–299. https://doi.org/10.1177/01466210022031796.
Coombs, D. H. (1969). Review of the analysis of essays by computer by Ellis B. Page and Dieter H. Paulus. Research in the Teaching of English 3: 222–228.
Cureton, E. F. (1951). Validity. In E. F. Lindquist (ed.), Educational Measurement, 1st edn. Washington, DC: American Council on Education, 621–694.
Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring [Special issue]. Language Testing 27 317–334. https://doi.org/10.1177/0265532210363144.
Fitzgerald, K. R. (1994). Computerized scoring? A question of theory and practice. Journal of Basic Writing 13: 3–17.
Jamieson, J. (2005). Trends in computer-based second language assessment. Annual Review of Applied Linguistics 24 228–242. https://doi.org/10.1017/S0267190505000127.
Kane, M. (1992). An argument-based approach to validity. Psychological Bulletin 112 527–535. https://doi.org/10.1037/0033-2909.112.3.527.
Kane, M. (2002). Validating high-stakes testing programs. Educational Measurement: Issues and Practice 21 31–35. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x.
Kane, M. (2006). Validation. In R. L. Brennan (ed.), Educational Measurement, 4th edn. Washington, DC: American Council on Education/Praeger, 18–64.
Kelly, P. A. (2001). Automated Scoring of Essays: Evaluating Score Validity. Unpublished dissertation, Florida State University.
Landauer, T. K. , Foltz, P. W. and Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes 25 259–284. https://doi.org/10.1080/01638539809545028.
Link, S. , Mehrzad, M. and Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2020.1743323.
Macrorie, K. (1969). Review of the analysis of essays by computer by Ellis B. Page and Dieter H. Paulus. Research in the Teaching of English 3: 222–236.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist 50 741–749. https://doi.org/10.1037/0003-066X.50.9.741.
Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. Educational Measurement: Issues and Practices 22 13–25. https://doi.org/10.1111/j.1745-3992.2003.tb00140.x.
Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. The Modern Language Journal 93 836–847. https://doi.org/10.1111/j.1540-4781.2009.00976.x.
Page, E. B. (1966). The imminence of grading essays by computers. Phi Delta Kappan 47: 238–243.
Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. Journal of Experimental Education 62 127–142. https://doi.org/10.1080/00220973.1994.9943835.
Page, E. B. and Dieter, P. (1995). The Analysis of Essays by Computer, Final Report of U.S. Office of Education Project No. 6–1318. Storrs, CT: University of Connecticut. ERIC ED 028 633.
Page, E. B. and Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. Phi Delta Kappan 76: 561–565.
Pearson . (2019). Pearson Test of English Academic: Automated Scoring. Retrieved January 2, 2021, from PTE Academic Automated Scoring White Paper (pearsonpte.com).
Petersen, N. S. (1997, March). Automated Scoring of Writing Essays: Can Such Scores Be Valid? Paper presented at The Annual Meeting of the National Council on Education, Chicago, IL.
Powers, D. E. , Bursterin, J. , Chodorow, M. S. , Fowles, M. E. and Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. Educational Computing Research 26 407–425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A.
Powers, D. E. , Burstein, J. C. , Fowles, M. E. and Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. Computer in Human Behavior 18 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8.
Roy, E. L. (1993). Computerized scoring of placement exams: A validation. Journal of Basic Writing 12 41–54. Retrieved December 21, 2020, from www.jstor.org/stable/43443612.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks 61 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
Shermis, M. D. , Koch, C. M. , Page, E. B. , Keith, T. Z. and Harrington, S. (1999, April). Trait Ratings for Automated Essay Grading. Paper presented at the Annual Meeting of National Council on Measurement in Education, Montreal, Canada.
Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. Language Testing 27 335–353. https://doi.org/10.1177/0265532210364406.
Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. Assessing Writing 18 85–99. https://doi.org/10.1016/j.asw.2012.10.006.
Xi, X. (2008). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In C. A. Chapelle , Y. R. Chung and J. Xu (eds.), Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment. Ames, IA/: Iowa State University, 102–114.
Xi, X. (2010). Automated scoring and feedback systems – Where are we and where are we heading? Language Testing 27 291–300. https://doi.org/10.1177/0265532210364643.
Xi, X. , Higgins, D. , Zechner, K. and Williamson, D. M. (2008). Automated Scoring of Spontaneous Speech Using SpeechRater v1.0 (ETS Research Rep. No. RR-08–62). Princeton, NJ: ETS.
Xi, X. , Schmidgall, J. and Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu and Y. Jin (eds.), Assessing Chinese Learners of English. London: Palgrave Macmillan.
Xi, X. and Zhang, C. Q. (2020). Validity and validation in language assessment: Developments and challenges. China Exams 6: 19–26.
Yang, Y. , Buckendahl, C. W. , Juszkiewicz, P. J. and Bhola, D. S. (2002). A review of strategies for validating computer automated scoring. Applied Measurement in Education 15 391–412. https://doi.org/10.1207/S15324818AME1504_04.
Zechner, K. and Evanini, K. (eds.). (2020). Automated speaking assessment: Using language technologies to score spontaneous speech. In J. Norris , S. Ross , S. Weigle and X. Xi (eds.), Innovations in Language Learning and Assessment at ETS. New York, NY: Routledge.
Zhang, Z. V. and Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. Assessing Writing 36 90–102. https://doi.org/10.1016/j.asw.2018.02.004.

Computer-based testing

Chapelle, C. A. and Douglas, D. (2006). Assessing Language Through Computer Technology. Cambridge, UK: Cambridge University Press. This Cambridge Language Assessment Series volume offers an excellent introduction to various aspects of CBT for students and professionals who are new to the topic. The authors start their discussion by identifying notable features of CBT as well as similarities and differences of CBT against conventional assessments. Various issues raised in chapters on the threat of CBT, evaluating CBT, and the impact of CBT are particularly eye opening. The authors' proposal of a context-sensitive definition of language ability for communication through technology marks an important step forward in enhancing our understanding of L2 abilities required to communicate successfully in technology-mediated environments.
Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (2008). Building a Validity Argument for the Test of English as a Foreign LanguageTM. New York, NY: Routledge. Chapelle, Enright, and Jamieson offer an overview of the history of the TOEFL as well as a comprehensive account of how the TOEFL iBT was designed and developed. Discussions on rationales behind various aspects of the test design and research studies conducted as part of the test development process and for initial validation of the test are valuable for the reader to understand the complexity and extensiveness of the work involved in the development of the high-volume, high-stakes CBT over the years. In addition, the TOEFL validity argument framework presented in the final chapter illustrates how an argument-based approach can be applied to CBT validation.
Chapelle, C. A. and Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. Language Learning & Technology 20: 116–128. This article is a review of papers and book reviews on CBT for language assessment published in the Language Learning & Technology journal between 1997 and 2015. Consistent with Bennet's (2000) three phases of CBT development adopted in this chapter, Chapelle and Voss discuss the studies identified according to two categories, technology for efficiency and technology for innovation. Based on the results, Chapelle and Voss note how deeply technology and language assessment are ingrained into language learning and emphasize the importance of teacher training for effective integration of technology and language assessment into the language classroom.
Litman, D. , Strik, H. and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. Language Assessment Quarterly 15: 294–309. https://doi.org/10.1080/15434303.2018.1472265. This paper is in the Language Assessment Quarterly

special issue, conceptualizing and operationalizing speaking assessment for a new century (guest edited by G. S. Lim), shows what automated speaking assessment might look like in the near future. The authors provide a non-technical explanation of how an automatic speech recognizer (ASR), which serves as the basis of automated speech scoring, could be combined with spoken dialogue systems (SDS) to emulate spoken interaction. The authors' discussion of the current state, challenges, and potentials of this new technology for L2 speaking assessment application is insightful, raising various important issues of further consideration concerning task design and construct representation in automated L2 speaking assessment.

Alderson, J. C. (2000a). Assessing Reading. Cambridge, UK: Cambridge University Press.

Alderson, J. C. (2000b). Technology in testing: The present and the future. System 28 593–603. https://doi.org/10.1016/S0346-251X(00)00040-3.

Alderson, J. C. (2005). Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment. New York, NY: Continuum.

Alderson, J. C. and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. Language Testing 22 301–320. https://doi.org/10.1191/0265532205lt310oa.

American Psychological Association . (1986). Guidelines for Computer-Based Tests and Interpretations. Washington, DC: Author.

Bachman, L. F. and Palmer, A. (1996). Language Testing in Practice. Cambridge, UK: Cambridge University Press.

Bachman, L. F. and Palmer, A. (2010). Language Assessment in Practice. Cambridge, UK: Cambridge University Press.

Bennett, R. E. (2000). Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing. Princeton, NJ: ETS.

Bernstein, J. , van Moere, A. and Cheng, J. (2010). Validating automated speaking tests. Language Testing 27 355–377. https://doi.org/10.1177/0265532210364404.

Blackhurst, A. (2005). Listening, reading, and writing on computer-based and paper-based versions of IELTS. University of Cambridge ESOL Examinations. Research Notes 21: 14–17.

Breland, H. , Lee, Y.-W. and Muraki, E. (2004). Comparability of TOEFL CBT Writing Prompts: Response Mode Analyses (TOEFL Research Report No. RR-75). Princeton, NJ: ETS. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01950.x.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. Language Learning & Technology 1 44–59. https://doi.org/10125/25003.

Brunfaut, T. , Harding, L. and Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. Assessing Writing 3 3–18. https://doi.org/10.1016/j.asw.2018.02.003.

Chalhoub-Deville, M. (ed.). (1999). Issues in Computer-adaptive Testing of Reading Proficiency. Cambridge, UK: Cambridge University Press.

Chalhoub-Deville, M. and Deville, C. (1999). Computer adaptive testing in second language contexts. Annual Review of Applied Linguistics 19 273–299. https://doi.org/10.1017/S0267190599190147.

Chapelle, C. A. and Douglas, D. (2006). Assessing Language Through Computer Technology Cambridge, UK: Cambridge University Press.

Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (2008). Building a Validity Argument for the Test of English as a Foreign Language. New York, NY: Routledge.

Chapelle, C. A. , Jamieson, J. and Hegelheimer, V. (2003). Validation of a web-based ESL test. Language Testing 20 409–439. https://doi.org/10.1191/0265532203lt266oa.

Chapelle, C. A. and Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. Language Learning & Technology 20 116–128. https://doi.org/10125/44464.

Choi, I.-C. , Kim, K. S. and Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. Language Testing 20 295–320. https://doi.org/10.1191/0265532203lt258oa.

Choi, I.-C. , Wolf, M. K. , Pooler, M. , Sova, L. and Faulkner-Bond, M. (2019). Investigating the benefits of scaffolding in assessments of young English learners: A case for scaffolded retell tasks. Language Assessment Quarterly 16 161–179. https://doi.org/10.1080/15434303.2019.1619180.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. Language Assessment Quarterly 3 295–306. https://doi.org/10.1207/s15434311laq0303_4.

Chun, C. W. (2008). Comments on 'Evaluation of the usefulness of the versant for English test: A response': The author responds. Language Assessment Quarterly 5 168–172. https://doi.org/10.1080/15434300801934751.

Chung, Y.-R. (2017). Validation of technology-assisted language tests. In C. A. Chapelle and S. Sauro (eds.), The Handbook of Technology and Second Language Teaching and Learning. New York, NY/: Wiley, 332–347.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English language listening test. ReCALL 18 193–211. https://doi.org/10.1017/S0958344006000425.
Council of Europe . (2001). Common European Framework of Reference for Languages: Learning Teaching and Assessment. Cambridge, UK: Cambridge University Press.
Douglas, D. (2013). Technology and language testing. In C. Chapelle (ed.), Encyclopedia of Applied Linguistics. New York, NY: Wiley. https://doi.org/10.1002/9781405198431.wbeal1182.
Douglas, D. and Hegelheimer, V. (2007). Assessing language using computer technology. Annual Review of Applied Linguistics 27 115–132. https://doi.org/10.1017/S0267190508070062.
Downey, R. , Farhady, H. , Present-Thomas, R. , Suzuki, M. and van Moere, A. (2008). Evaluation of the usefulness of the versant for English test: A response. Language Assessment Quarterly 5 160–167. https://doi.org/10.1080/15434300801934744.
Fulcher, G. (1999). Computerizing an English language placement test. ELT Journal 53: 289–298.
Fulcher, G. (2003). Interface design in computer-based language testing. Language Testing 20 384–408. https://doi.org/10.1191/0265532203lt265oa.
Galaczi, E. and Taylor, L. (2018). Interactional Competence: Conceptualisations, operationalisations, and outstanding questions. Language Assessment Quarterly 15 219–236. https://doi.org/10.1080/15434303.2018.1453816.
Jamieson, J. (2005). Trends in computer-based second language assessment. Annual Review of Applied Linguistics 25 228–242. https://doi.org/10.1017/S0267190505000127.
Jamieson, J. , Eignor, D. , Grabe, W. and Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. Chapelle , M. K. Enright and J. Jamieson (eds.), Building a Validity Argument for the Test of English as a Foreign LanguageTM. New York, NY: Routledge.
Jin, Y. and Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. Language Assessment Quarterly 14 101–119. https://doi.org/10.1080/15434303.2016.1261293.
Kane, M. (2006). Validation. In R. Brennan (ed.), Educational Measurement, 2nd edn. Westport, CT: American Council on Education and Praeger Publishers.
Kenyon, D. M. and Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. Language Learning Technology 5 60–83. https://doi.org/10125/25128.
Kunnan, A. and Carr, N. (2017). A comparability study between the general English proficiency test-advanced and the internet-based test of English as a Foreign language. Language Testing in Asia 7: 17. https://doi.org/10.1186/s40468-017-0048-x.
Lee, J.Y. (2002). A Comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. Assessing Writing 8 135–157. https://doi.org/10.1016/S1075-2935(03)00003-5.
Lee, Y.-W. and Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. Language Assessment Quarterly 6 239–263. https://doi.org/10.1080/15434300903079562.
Ling, G. (2017). Is writing performance related to keyboard type? An investigation from examinees' perspectives on the TOEFL iBT. Language Assessment Quarterly 14 36–53. https://doi.org/10.1080/15434303.2016.1262376.
Litman, D. , Strik, H. and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. Language Assessment Quarterly 13 294–309. https://doi.org/10.1080/15434303.2018.1472265.
Malabonga, V. , Kenyon, D. M. and Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. Language Testing 22 59–92. https://doi.org/10.1191/0265532205lt297oa.
Milton, J. (2013). Second language acquisition via Second Life. In C. Chapelle (ed.), Encyclopedia of Applied Linguistics. New York, NY: Wiley. https://doi.org/10.1002/9781405198431.wbeal1318.
Nakatsuhara, F. , Inoue, C. , Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. Language Assessment Quarterly 14 1–18. https://doi.org/10.1080/15434303.2016.1263637.
Nogami, Y. and Hayashi, N. (2010). A Japanese adaptive test of English as a foreign language: Developmental and operational aspects. In W. van and der Linden C. A. W. Glas (eds.), Elements of Adaptive Testing. New York, NY: Springer.
Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment. Language Learning and Technology 5 99–105. https://doi.org/10125/25131.
Ockey, G. J. , Gu, L. and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. Language Assessment Quarterly 14 346–359. https://doi.org/10.1080/15434303.2017.1400036.

Ockey, G. J. , Timple-Loughlin, V. , Davis, L. and Gu, L. (2019). Exploring the Potential of a Video-Mediated Interactive Speaking Assessment. (ETS Research Report No. ETS RR-19–05). Princeton, NJ: ETS. https://doi.org/10.1002/ets2.12240.
Ockey, G. J. and Wagner, E. (2018). Assessing L2 Listening: Moving Towards Authenticity. Amsterdam, Netherlands: John Benjamins.
Park, M. (2018). Innovative assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. ReCALL 30 196–213. https://doi.org/10.1017/S0958344017000362.
Plough, I. , Banerjee, J. and Iwashita, N. (2018). Interactional competence: Genie out of the bottle. Language Testing 35 427–445. https://doi.org/10.1177/0265532218772325.
Poehner, M. E. (2008). Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development. Berlin: Springer.
Poehner, M. E. and Lantolf, J. P. (2013). Bridging the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). Language Teaching Research 17 323–342. https://doi.org/10.1177/1362168813482935.
Poehner, M. E. , Zhang, J. and Lu, X. (2014). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. Language Testing 32 337–357. https://doi.org/10.1177/0265532214560390.
Roever, C. (2001). Web-based language testing. Language Learning and Technology 5 84–94. https://doi.org/10125/25129.
Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. Language Learning and Technology 5 38–59. https://doi.org/10125/25127.
Schmidgall, J. E. (2017). Articulating and Evaluating Validity Arguments for the TOEIC® Tests. (ETS Research Report No. ETS RR-17–51). Princeton, NJ: ETS. https://doi.org/10.1002/ets2.12182.
Schmidgall, J. E. and Powers, D. E. (2017). Technology and high-stakes language testing. In C. A. Chapelle and S. Sauro (eds.), The Handbook of Technology and Second Language Teaching and Learning. New York, NY/: Wiley, 317–331.
Schroeders, U. and Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. Educational and Psychological Measurement 71 849–869. https://doi.org/10.1177/0013164410391468.
Stansfield, C. W. (ed.). (1986). Technology and Language Testing. Washington, DC: TESOL.
Taylor, C. , Jamieson, J. , Eignor, D. and Kirsch, I. (1998). The Relationship Between Computer Familiarity and Performance on Computer-Based TOEFL Test Tasks (TOEFL Research Report No. 61). Princeton, NJ: ETS. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01757.x.
Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. Language Learning & Technology 16 10–20. https://doi.org/10125/44292.
Trites, L. and McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? Language Testing 22 174–210. https://doi.org/10.1191/0265532205lt299oa.
Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (eds.), Handbook of Second Language Assessment. Berlin/: De Gruyter, 255–273.
Vygotsky, L. S. (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. Basingstoke, UK: Palgrave Macmillan.
Wolf, M. K. , Guzman-Orth, D. , Lopez, A. , Castellano, K. , Himelfarb, I. and Tsutagawa, F. S. (2016). Integrating scaffolding strategies into technology enhanced Assessments of English learners: Task types and measurement models. Educational Assessment 21 157–175. https://doi.org/10.1080/10627197.2016.1202107.
Wolfe, E. W. and Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. Language Learning & Technology 8 53–65. http://doi.org/10125/25229.
Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. Language Assessment Quarterly 7 119–136. https://doi.org/10.1080/15434300903452355.

Corpus linguistics and language testing

Cushing, S. T. (2017). Corpus linguistics in language testing research [special issue]. Language Testing 34: 441–449. This issue of *Language Testing* contains five original research articles that illustrate different applications of corpus linguistics data and analysis tools to language assessment, along with an introduction and two commentaries: one by a corpus linguist and the other by an assessment specialist.

Gries, S. T. (2010). Useful statistics for corpus linguistics. A Mosaic of Corpus Linguistics: Selected Approaches 66: 269–291. This article provides an overview of statistics that are useful for analyzing corpus data, which consists essentially of distributional frequency data.
O'Keeffe, A. and McCarthy, M. (eds.). (2010). The Routledge Handbook of Corpus Linguistics. London: Routledge. This edited volume of 45 chapters provides a good introduction to the field of corpus linguistics, with sections devoted to corpus compilation, corpus analysis, and the use of corpora for language analysis, education, and other applications.
Adolphs, S. , Brown, B. , Carter, R. , Crawford, P. and Sahota, O. (2007). Applying corpus linguistics in a health care context. Journal of Applied Linguistics and Professional Practice 1 9–28. https://doi.org/10.1558/japl.v1.i1.9.
Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas and M. Short (eds.), Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech. London/: Longman, 248–259.
Alderson, J. C. and Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. Language Testing 30 535–556. https://doi.org/10.1177/0265532213489568.
Anthony, L. (2012). AntConc [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from www.antlab.sci.waseda.ac.jp/.
Anthony, L. (2013). A critical look at software tools in corpus linguistics. Linguistic Research 30 141–161. https://doi.org/10.17250/khisli.30.2.201308.001.
Aull, L. (2017). Corpus analysis of argumentative versus explanatory discourse in writing task genres. Journal of Writing Analytics 1: 1.
Austen, J. (1994). Pride and Prejudice. 1813. Retrieved from www.pemberley.com/janeinfo/pridprej.html.
Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quarterly, 2 1–34. https://doi.org/10.1207/s15434311laq0201_1.
Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.
Bachman, L. F. and Palmer, A. S. (2010). Language Assessment in Practice. Oxford: Oxford University Press.
Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (ed.), Research Methods in Linguistics. London/: Bloomsbury, 93–113.
Barker, F. (2006). Corpora and language assessment: Trends and prospects. Research Notes 26: 2–4.
Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe and M. McCarthy (eds.), The Routledge Handbook of Corpus Linguistics. Abingdon, UK/: Routledge, 661–674.
Barker, F. (2013). Using corpora to design assessment. In A. J. Kunnan (ed.), The Companion to Language Assessment. Hoboken, NJ/: Wiley-Blackwell, 1013–1028.
Barlow, M. (2012). MonoConc Pro 2.2 (MP2.2) [Software].
Baron, A. , Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. Anglistik 20: 41–67.
Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing 8: 243–257.
Biber, D. and Conrad, S. (2009). Register, Genre, and Style. Cambridge, UK: Cambridge University Press.
Biber, D. , Conrad, S. and Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. Applied linguistics 25 371–405. https://doi.org/10.1093/applin/25.3.371.
Biber, D. , Conrad, S. and Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge, UK: Cambridge University Press.
Biber, D. , Conrad, S. , Reppen, R. , Byrd, P. , Helt, M. , . Clark, V. … Urzua, A. (2004). Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. TOEFL Monograph Series MS-25. Princeton, NJ: Educational Testing Service.
Callies, M. and Götz, S. (eds.). (2015). Learner Corpora in Language Testing and Assessment, vol. 70. Amsterdam, The Netherlands: John Benjamins Publishing Company.
Chapelle, C. A. , Enright, M. K. and Jamieson, J. M. (eds.). (2008). Building a Validity Argument for the Test of English as a Foreign LanguageTM. Abingdon, UK: Routledge.
Chapelle, C. A. and Plakans, L. (2013). Assessment and testing: Overview. In C. A. Chapelle (ed.), The Encyclopedia of Applied Linguistics. Hoboken, NJ/: Wiley-Blackwell, 241–242.
Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. International Journal of Corpus Linguistics 18 418–442. https://doi.org/10.1075/ijcl.18.3.07che.
Chujo, K. and Utiyama, M. (2006). Selecting level-specific vocabulary using statistical measures. System 34 255–269. https://doi.org/10.1016/j.system.2005.12.003.
Cobb, T. (2019). Compleat Lexical Tutor. Retrieved December 19, 2019, from www.lextutor.ca/conc/eng/.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. Calico Journal, 14(2–4), 15–33. Retrieved from www.jstor.org/stable/45119223.
Corpus Resource Database (CoRD). Department of English, University of Helsinki . Retrieved December 15, 2019, from www.helsinki.fi/varieng/CoRD/corpora/.
Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching, assessment. Cambridge: Cambridge University Press.
Cushing, S. T. (2017). Corpus linguistics in language testing research [special issue]. Language Testing 34: 441–449.
Crosthwaite, P. R. and Raquel, M. (2019). Validating an L2 academic group oral assessment: Insights from a spoken learner corpus. Language Assessment Quarterly 16 1–25. https://doi.org/10.1080/15434303.2019.1572149.
Dagneaux, E. , Denness, S. and Granger, S. (1998). Computer-aided erroranalysis. System 26: 163–174.
Davies, M. (2008–). The Corpus of Contemporary American English (COCA): 520 Million Words, 1990-Present. Retrieved from http://corpus.byu.edu/coca/.
Davies, M. (2018). The iWeb Corpus: 14 Billion Words. Retrieved December 28, 2019.
Davis, J. M. , Norris, J. M. , Malone, M. E. , , McKayT. H. and Son, Y. (2018). Useful Assessment and Evaluation in Language Education. Washington, DC: Georgetown University Press. Retrieved from muse.jhu.edu/book/59038.
Díaz-Negrillo, A. and Domínguez, J. F. (2006). Error tagging systems for learner corpora. Revista española de lingüística aplicada, 83–102.
Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. Language Testing 27 317–334. https://doi.org/10.1177/0265532210363144.
Francis, W. N. and Kucera, H. (1979). Brown Corpus Manual: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for Use with Digital Computers. Providence, RI: Brown University.
Friginal, E. (2009). The Language of Outsourced Call Centers: A Corpus-Based Study of Cross-Cultural Interaction, vol. 34. Amsterdam, The Netherlands: John Benjamins Publishing.
Friginal, E. and Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. Journal of Second Language Writing 26: 80–95.
Gablasova, D. (2020). Corpora for second language assessments. In P. Winke and T. Brunfaut (eds.), The Routledge Handbook of Second Language Acquisition and Language Testing. Abingdon/: Routledge, 45–53.
Gablasova, D. , Brezina, V. and McEnery, T. (2019). The trinity Lancaster corpus: Development, description and application. International Journal of Learner Corpus Research 5 126–158. https://doi.org/10.1075/ijlcr.19001.gab.
Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside , G. Leech and G. Sampson (eds.), The Computational Analysis of English: A Corpus-Based Approach. London: Longman.
Graesser, A. C. , McNamara, D. S. , Louwerse, M. M. and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments and Computers 36 193–202. https://doi.org/10.3758/BF03195564.
Granger, S. , Dagneaux, E. , Meunier, F. and Paquot, M. (2009). The International Corpus of Learner English. Handbook and CD-ROM. Version 2. Louvain-la-Neuve: Presses universitaires de Louvain.
Gries, S. T. (2009). What is corpus linguistics? Language and Linguistics Compass 3 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x.
Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sanchez and M. Almea (eds.), A Mosaic of Corpus Linguistics: Selected Approaches. Frankfurt am Main/: Peter Lang, 269–291.
Hawkey, R. and Barker, F. (2004). Developing a common scale for the assessment of writing. Assessing Writing 9: 122–159.
Hawkins, J. A. and Filipović, L. (2012). Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework, vol. 1. Cambridge, UK: Cambridge University Press.
Kane, M. (2006). Validation. In R. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: American Council on Education and Praeger, 17–64.
Kane, M. (2012). Validating score interpretations and uses. Language Testing 29 3–17. https://doi.org/10.1177/0265532211417210.
Kennedy, G. (1998). An Introduction to Corpus Linguistics. London: Longman.
Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. TESOL Quarterly 37: 467–487.
Kurtes, S. and Saville, N. (2008). The English profile programme – An overview. Research Notes 33: 2–4.
LaFlair, G. T. and Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. Language Testing 34 451–475.

https://doi.org/10.1177/0265532217713951.
Latifi, S. and Gierl, M. (2020). Automated scoring of junior and senior high essays using Coh- Metrix features: Implications for large-scale language testing. Language Testing 38 62–85. https://doi.org/10.1177/0265532220929918.
Learner Corpora Around the World . Retrieved December 15, 2019, from https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.
Leech, G. , Garside, R. and Bryant, M. (1994). Corpus-based research into language: In honour of Jan Aarts. In N. Oostdjik and P. Haan (eds.), The Large-Scale Grammatical Tagging of Text: Experience with the British National Corpus. Netherlands/: Rodopi Publishers, 47–63.
Love, R. , Dembry, C. , Hardie, A. , Brezina, V. and McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics 22 319–344. https://doi.org/10.1075/ijcl.22.3.02lov.
Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics 15 474–496. https://doi.org/10.1075/ijcl.15.4.02lu.
Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Language Testing 34 493–511. https://doi.org/10.1177/0265532217710675.
Lüdeling, A. , Walter, M. , Kroymann, E. and Adolphs, P. (2005). Multi-level error annotation in learner corpora. Proceedings of Corpus Linguistics 2005: 14–17.
McArthur, T. (1981). Longman Lexicon of Contemporary English. London: Longman.
McEnery, T. and Wilson, A. (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press.
Michigan Corpus of Academic Spoken English. (2009). Ann Arbor, MI: The Regents of the University of Michigan.
Michigan Corpus of Upper-level Student Papers. (2009). Ann Arbor, MI: The Regents of the University of Michigan.
Moder, C. L. and Halleck, G. B. (2012). Designing language tests for specific social uses. In Routledge Handbook of Language Testing. Abingdon/: Routledge, 137–189.
Nesi, H. (2001). A corpus-based analysis of academic lectures across disciplines. In J. Cotterill and A. Ife (eds.), Language Across Boundaries. London/: Continuum, 201–218.
Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. Language Assessment Quarterly 15 29–43. https://doi.org/10.1080/15434303.2017.1405421.
Park, K. (2014). Corpora and language assessment: The state of the art. Language Assessment Quarterly 11 27–44. https://doi.org/10.1080/15434303.2013.872647.
Plakans, L. (2018). Then and now: Themes in language assessment research. Language Education & Assessment 1 3–8. https://doi.org/10.29140/lea.v1n1.35.
Rayson, P. , Archer, D. , Piao, S. and McEnery, A. M. (2004). The UCREL Semantic Analysis System. Lancaster, UK: Lancaster University.
Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. International Journal of English Studies 7: 105–126.
Reppen, R. and Simpson, R. (2002). Corpus linguistics. In N. Schmitt (ed.), An Introduction to Applied Linguistics. London/: Arnold Publishers, 92–111.
Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. Language Testing 34 477–492. https://doi.org/10.1177/0265532217711431.
Scott, M. (2016). WordSmith Tools Version 7. Stroud: Lexical Analysis Software.
Simpson, R. C. , Lee, D. W. and Leicher, S. (2002). MICASE Manual. The Michigan Corpus of Academic English. Ann Arbor, MI: The University of Michigan.
Stubbs, M. and Halbe, D. (2012). Corpus linguistics: Overview. In C. Chapelle (ed.), The Encyclopedia of Applied Linguistics. Hoboken, NJ: John Wiley & Sons.
Subirats, C. and Ortega, M. (2012). Corpus del Español Actual. Retrieved from http://spanishfn.org/tools/cea/english.
Svartvik, J. (2011). Corpus Linguistics Comes of Age. Directions in corpus linguistics: Proceedings of Nobel symposium 82 Stockholm, 4–8 August 1991, 7–13.
Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. The Modern Language Journal 97 77–101. https://doi.org/10.1111/j.1540-4781.2012.01422.x.
Tribble, C. (2012, July 11–14). Teaching and Language Corpora: Quo Vadis? 10 th Teaching and Language Corpora Conference (TALC), Warsaw.
Van Moere, A. , Suzuki, M. , Downey, R. and Cheng, J. (2009). Implementing ICAO language proficiency requirements in the versant aviation English test. Australian Review of Applied Linguistics 32 27–29. https://doi.org/10.2104/aral0927.

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. Graduate Theses and Dissertations. 12691. https://lib.dr.iastate.edu/etd/12691.
Weigle, S. C. and Goodwin, S. (2016). Applications of corpus linguistics in language assessment. In J. Banerjee and D. Tsagari (eds.), Contemporary Second Language Assessment. New York, NY/: Bloomsbury, 209–224.
Xi, X. (2017). What does corpus linguistics have to offer to language assessment? Language Testing 34 565–577. https://doi.org/10.1177/0265532217720956.

Ethics and fairness

Kunnan, A. J. (2008). Using the test fairness and wider context frameworks. In L. Taylor and C. Weir (eds.), Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity. Papers from the ALTE Berlin Conference. Cambridge, UK: Cambridge University Press. This is a presentation of Kunnan's Test Context Framework, giving an example of a "macro-analytic" approach to investigating test fairness.
Lee, Y. (2010). Identifying suspect item bundles for the detection of differential bundle functioning in an ESL reading comprehension test: A preliminary study. In A. J. Kunnan (ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing 9. Cambridge, UK: Cambridge University Press, 105–127. This paper presents an example of a "micro-analytic" fairness study, taking concepts and techniques of differential item functioning (DIF) and applying them to the "test bundle": i.e., a group of related items on a test.
Toulmin, S. , Rieke, R. and Janik, A. (1984). An Introduction to Reasoning, 2nd edn. New York, NY: Macmillan. This is a good resource for understanding the formal antecedents of Xi's (2010) embedded fairness/validity argument structure.
Willingham, W. W. (1999). A systemic view of test fairness. In S. Messick (ed.), Assessment in Higher Education: Issues in Access, Quality, Student Development, and Public Policy. Mahwah, NJ: Lawrence Erlbaum Associates, 213–242. This book chapter is a good, short exploration of fairness issues at all stages of the test development process, from test design and item writing to test administration, interpretation, and use.
Yan, X. , Cheng, L. and Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examine L1 background on scores of an ITA speaking test. Language Testing 36: 207–234. This is a good example of a study that implements Xi's (2010) approach to investigating fairness. It first examines, via confirmatory factor analysis (CFA), the factor structure of responses to three different integrated tasks across three different L1 groups and then interprets any bias in the context of the assessment of international teaching assistants (ITAs), shedding light on construct-related evidence of the validity of the tasks, as well as fairness. While the study is clearly "micro-analytic," the authors' wider considerations of "radically different language backgrounds and experiences" of test takers suggest the possibility of a synthesis of "macro" and "micro" approaches to fairness.
AERA/APA/NCME (American Educational Research Association, American Psychological Association and National Council on Measurement in Education) . (1999). Standards for Educational and Psychological Testing. Washington, DC: Author.
Ayer, A. J. (1936). Language, Truth, and Logic. London: Penguin Modern Classics. Cited In G. Fulcher and F. Davidson (eds.). (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York, NY: Routledge.
Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford, UK: Oxford University Press.
Camilli, G. (2006). Test fairness. In R. Brennan (ed.), Educational Measurement, 4th edn. Westport, CT: American Council on Education and Praeger, 221–256.
Code of Fair Testing Practices in Education . (2004). The Code of Fair Testing Practices in Education. Washington, DC: Joint Committee on Testing Practices. Retrieved July 18, 2020, from www.apa.org/science/programs/testing/fair-testing.pdf.
Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin 52 281–302. https://doi.org/10.1037/h0040957.
Davies, A. (1997). Demands of being professional in language testing. Language Testing 14 328–339. https://doi.org/10.1177/026553229701400309.
Davies, A. (2010). Test fairness: A response. Language Testing 27: 171–176.
Dorans, N. J. and Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. Journal of Educational Measurement 23 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x.

Fulcher, G. (2012). Assessment literacy for the language classroom. Language Assessment Quarterly 9 113–132. https://doi.org/10.1080/15434303.2011.642041.
Fulcher, G. and Davidson, F. (eds.). (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York, NY: Routledge.
Hamp-Lyons, L. (1989). Language testing and ethics. Prospect 5: 7–15.
Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. Language Testing, 14 295–303. https://doi.org/10.1177/026553229701400306.
Hamp-Lyons, L. (2000). Social, professional, and individual responsibility in language testing. System 28(4) 579–598. https://doi.org/10.1016/S0346-251X(00)00039-7.
Hatcher, W. S. (2004). Minimalism. Hong Kong: Juxta Publishing.
Holland, P. W. and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (eds.), Test Validity. Hillsdale, NJ/: Lawrence Erlbaum Associates, 129–145.
Honer, S. M. , Hunt, T. C. and Okholm, D. L. (eds.). (2002). Invitation to Philosophy: Issues and Opinions. Belmont, CA: Wadsworth/Thomson Learning.
Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. Language Testing 25 385–402. https://doi.org/10.1177/0265532208090158.
International Language Testers Association . (2018). Code of Ethics. Retrieved July 27, 2020, from www.iltaonline.com/page/CodeofEthics.
Jang, E. (2002). Folk Fairness in Language Testing. Paper presented at the Southern California association for language assessment research conference (SCALAR 5), May 15–16.
Kane, M. (2010). Validity and fairness. Language Testing 27 177–182. https://doi.org/10.1177/0265532209349467.
Kant, E. (1999 [1788]). Critique of Practical Reason (M. J. Gregor , trans.). Cambridge, UK: Cambridge University Press.
Klemke, E. D. , Hollinger, R. and Rudge, D. W. (eds.). (1998). Introductory Readings in the Philosophy of Science, 3rd edn. Amherst, NY: Prometheus Books.
Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing 9. Cambridge, UK/: Cambridge University Press, 1–14.
Kunnan, A. J. (2004). Test fairness. In M. Milanovic and C. Weir (eds.), European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference. Cambridge, UK: Cambridge University Press.
Kunnan, A. J. (2009). Testing for citizenship: The U.S. Naturalization test. Language Assessment Quarterly 6 89–97. https://doi.org/10.1080/15434300802606630.
Lynch, B. K. (1997). In search of the ethical test. Language Testing 14 315–327. https://doi.org/10.1177/026553229701400308.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 955–966.
Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 1012–1027.
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum.
Messick, S. (1989). Validity. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: American Council on Education and Macmillan, 13–103.
Mish, F. C. (ed.). (2003). Merriam-Webster's Collegiate Dictionary, 11th edn. Springfield, MA: Merriam-Webster, Incorporated.
Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland and H. Wainer (eds.), Differential Item Functioning. Hillsdale, NJ/: Lawrence Erlbaum, 367–388.
Schwandt, T. A. and Jang, E. E. (2002). Linking validity and ethics in language testing: Insights from the hermeneutic turn in social science. Studies in Educational Evaluation 30 265–280. https://doi.org/10.1016/j.stueduc.2004.11.001.
Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (ed.), Fairness and Validation in Language Assessment. Cambridge, UK/: Cambridge University Press, 15–19.
Shohamy, E. (2001). The Power of Tests: A Critical Perspective on the Uses of Language Tests. Harlow, UK: Longman.
Spaan, M. (2000). Enhancing fairness through a social contract. In A. J. Kunnan (ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing 9. Cambridge, UK/: Cambridge University Press, 35–37.
Walter, E. (ed.). (2006). Cambridge Advanced Learner's Dictionary, 3rd edn. Cambridge: Cambridge University Press. Retrieved July 18, 2020, from dictionary.cambridge.org/dictionary/british/fair_1.

Willingham, W. W. and Cole, N. (1997). Gender and Fair Assessment. Mahwah, NJ: Lawrence Erlbaum Associates.
Xi, X. (2010). How do we go about investigating test fairness? Language Testing 27 147–170. https://doi.org/10.1177/0265532209349465.
Yan, X. , Cheng, L. and Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examine L1 background on scores of an ITA speaking test. Language Testing, 36 207–234. https://doi.org/10.1177/0265532218775764.
YourDictionary.com . (2019). YourDictionary.com. Burlingame, CA: LoveToKnow Corporation. Retrieved July 27, 2020, from www.yourdictionary.com/fairness.

Standards in language proficiency measurement

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ: Educational Testing Service, 1–13. When tracing the history of language performance standards, it is worth going back to first-hand accounts of the development of the precursors to today's standards.
Stein, Z. (2016). Social Justice and Educational Measurement. Oxon and New York, NY: Routledge. This original and insightful book applies ideas from moral and political philosophy to educational assessment. Stein argues that standardized assessment has not led to increased educational equality and proposes a model that offers a solution for how standardized testing could be used to foster equal educational opportunities.
Trim, J. L. M. (2012). The common European framework of reference for languages and its background: A case study of cultural politics and educational influences. In M. Byram and L. Parmenter (eds.), The Common European Framework of Reference: The Globalisation of Language Education Policy. Buffalo, NY: Multilingual Matters, 14–36. John Trim, one of the people who laid the groundwork for the CEFR, remembers which ideas, intentions, and aspirations inspired the creation of what was to become perhaps the most widely used set of language performance standards.
ACTFL . (2012). ACTFL Proficiency Guidelines 2012. Washington, DC: American Council on the Teaching of Foreign Languages.
ACTFL . (2016). Assigning CEFR Ratings to ACTFL Assessments. American Council on the Teaching of Foreign Languages. Retrieved from www.actfl.org/publications/guidelines-and-manuals/assigning-cefr-ratings-actfl-assessments.
Alderson, J. C. (2007). The CEFR and the need for more research. The Modern Language Journal 91 659–663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x.
Bachman, L. and Palmer, A. (2010). Language Assessment in Practice. Oxford: Oxford University Press.
Baker, B. A. (2014). Investigating Language Assessment Literacy in Canadian University Admissions. Amsterdam: LTRC.
Barni, M. (2015). In the name of the CEFR: Individuals and standards. In B. Spolsky , O. Inbar-Lourie and M. Tannenbaum (eds.), Challenges of Language Education and Policy. Making Space for People. New York/: Routledge, 40–52.
British Council . (2019a). IELTS grows to 3.5 million a year. Take IELTS. Retrieved from https://takeielts.britishcouncil.org/about/press/ielts-grows-three-half-million-year.
British Council . (2019b). Understand and explain the IELTS scores. Take IELTS. Retrieved from https://takeielts.britishcouncil.org/teach-ielts/test-information/scores-explained.
Brooks, R. L. and Hoffman, M. (2013). Government and military assessment. In The Companion to Language Assessment. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla069/abstract.
Cantwell, B. (2015). Are international students cash cows? Examining the relationship between new international undergraduate enrollments and institutional revenue at public colleges and universities in the US. Journal of International Students 5: 512–525.
Carroll, J. B. (1954). Notes on the Measurement of Achievement in Foreign Languages.
Centre for Canadian Language Benchmarks . (2012). Canadian Language Benchmarks. English as a Second Language for Adults. Ottawa, ON, Canda: Citizenship and Immigration Canada.
Council of Europe . (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Strasbourg, France: Council of Europe.
Council of Europe . (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Council of Europe Language Policy Division.

Cox, T. L. , Malone, M. E. and Winke, P. (2018). Future directions in assessment: Influences of standards and implications for language learning. Foreign Language Annals 51 104–115. https://doi.org/10.1111/flan.12326.
Davies, A. (2008). Assessing Academic English: Testing English Proficiency 1950–89: The IELTS Solution. Cambdridge, UK: Cambridge University Press.
De Jong, J. H. A. L., Becker, K. , Bolt, D. and Goodman, J.Pearson (2014). Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages. . Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf.
Deygers, B. (2019). The CEFR companion volume: Between research-based policy and policy-based research. Applied Linguistics. https://doi.org/10.1093/applin/amz024.
Deygers, B. and Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. Language Testing 36 347–368. https://doi.org/10.1177/0265532219826390.
Deygers, B. , Zeidler, B. , Vilcu, D. and Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. Language Assessment Quarterly 15 3–15. https://doi.org/10.1080/15434303.2016.1261350.
Dubeau, J. (2006). Are We All on the Same Page? An Exploratory Study of OPI Ratings Across NATO Countries Using the NATO STANAG 6001 Scale. Ottawa, ON, Canada: Carleton University.
ETS . (2010). Linking TOEFL iBT TM Scores to IELTS® Scores – A Research Report. Princeton, NJ: Educational Testing Service.
ETS . (2018). Test and Score Data Summary for TOEFL iBT® Tests. January 2017 – December 2017. Princeton, NJ: Educational Testing Service.
ETS . (2019). Performance Descriptors for the TOEFL iBT® Test. Princeton, NJ: Educational Testing Service.
Figueras, N. (2012). The impact of the CEFR. ELT Journal 66 477–485. https://doi.org/10.1093/elt/ccs037.
Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. System 24: 163–172.
Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. Language Assessment Quarterly 1 253–266. https://doi.org/10.1207/s15434311laq0104_4.
Fulcher, G. (2012). Scoring performance tests. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London and New York/: Routledge, 378–392.
Fulcher, G. (2015). Re-examining Language Testing: A Philosophical and Social Inquiry. London and New York: Routledge.
Fulcher, G. (2016). Standards and frameworks. In D. Tsagari and J. Banerjee (eds.), Handbook of Second Language Assessment. Berlin/: De Gruyter Mouton, 29–44.
Fulcher, G. , Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. Language Testing 28 5–29. https://doi.org/10.1177/0265532209359514.
Gaber, S. , Cankar, G. , Umek, L. M. and Tašner, V. (2012). The danger of inadequate conceptualisation in PISA for education policy. Compare: A Journal of Comparative and International Education 42 647–663. https://doi.org/10.1080/03057925.2012.658275.
Galaczi, E. D. , ffrench, A. , Hubbard, C. and Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. Assessment in Education: Principles, Policy & Practice 18 217–237. https://doi.org/10.1080/0969594X.2011.574605.
Glisan, E. W. (2012). National standards: Research into practice. Language Teaching 45 515–526. https://doi.org/10.1017/S0261444812000249.
Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. Language Assessment Quarterly 15 59–74. https://doi.org/10.1080/15434303.2017.1350685.
Green, R. and Wall, D. (2005). Language testing in the military: Problems, politics and progress. Language Testing 22 379–398. https://doi.org/10.1191/0265532205lt314oa.
The Guardian . (2014). OECD and Pisa tests are damaging education worldwide – Academics. The Guardian. Retrieved from www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics.
Hamid, M. O. and Hoang, N. T. H. (2019). Humanising language testing. TSL-EJ 22.
Harvard University . (2019). International Applicants. Harvard College. Retrieved from https://college.harvard.edu/admissions/apply/international-applicants.
Hatto, P. (2010a). Standards and Standardisation. A Practical Guide for Researchers. Brussels: European Commission.
Hatto, P. (2010b). Standards and Standardization Handbook. Brussels: European Commission.
Hudson, T. (2012). Standards-based testing. In G. Fulcher and F. Davidson (eds.), The Routledge Handbook of Language Testing. London and New York/: Routledge, 479–495.
Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. The Modern Language Journal 91 663–667. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x.

Hyatt, D. (2013). Stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. Journal of Further and Higher Education 37 844–863. https://doi.org/10.1080/0309877X.2012.684043.
Hyatt, D. and Brooks, G. (2009). Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. IELTS Research Reports 10, 17–68.
IEA . (2019). About IEA. Retrieved from www.iea.nl/.
IELTS . (2019). About IELTS USA. IELTS. Retrieved from www.ielts.org/usa/about-ielts-usa.
International Organization for Standardization . (2016). How to Write Standards. Geneva: International Organization for Standardization.
International Organization for Standardization . (2018). International Standards & Trade Agreements. Geneva: World Standards Cooperation.
Jin, Y. , Wu, Z. , Alderson, C. and Song, W. (2017). Developing the China standards of English: Challenges at macropolitical and micropolitical levels. Language Testing in Asia 7: 1. https://doi.org/10.1186/s40468-017-0032-5.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement 50 1–73. https://doi.org/10.1111/jedm.12000.
Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. The Modern Language Journal 28 136–150. https://doi.org/10.2307/317331.
JSTOR.Krumm, H-J. . (2007). Profiles instead of levels: The CEFR and its (Ab)uses in the context of migration. The Modern Language Journal 91 667–669. https://doi.org/10.1111/j.1540-4781.2007.00627_6.x.
Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: Gateway to testing and curriculum. Foreign Language Annals 17 475–489. http://dx.doi.org.kuleuven.ezproxy.kuleuven.be/10.1111/j.1944-9720.1984.tb01736.x.
Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. Foreign Language Annals 36 483–490. https://doi.org/10.1111/j.1944-9720.2003.tb02137.x.
Liss, J. M. (2013). Creative destruction and globalization: The rise of massive standardized education platforms. Globalizations 10 557–570. https://doi.org/10.1080/14747731.2013.806741.
Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. The Modern Language Journal 91 645–655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x.
Little, D. (2019). Proficiency guidelines and frameworks. In J. Schwieter and A. Benati (eds.), The Cambridge Handbook of Language Learning. Cambdrige, UK/: Cambridge University Press, 550–574.
Massachusetts Institute of Technology . (2019). IELTS Requirement for International Students. MIT Media Lab. Retrieved from www.media.mit.edu/posts/ielts-requirement-for-international-students/.
McNamara, T. (2014). 30 years on – Evolution or revolution? Language Assessment Quarterly 11 226–232. https://doi.org/10.1080/15434303.2014.895830.
Ministry of Education of the People's Republic of China . (2018). National Language Standard. China's Standards of English Language Ability. (GF 0018–2018). Ministry of Education of the People's Republic of China and National Language Commission of the People's Republic of China.
Mullis, I. V. S. , Martin, M. O. , Foy, P. and Hooper, M. (2017). PIRLS 2016. International Results in Reading. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement.
NATO . (2014). Standardization Agreement STANAG 6001 Language Proficiency Levels. Brussels: Bureau for International Language Coordination.
Negishi, M. and Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. In C. Docherty and F. Barker (eds.), Language Assessment for Multilingualism. Proceedings of the ALTE Paris Conference, April 2014. Cambridge/: Cambridge University Press, 113–134.
North, B. (2014a). English Profile Studies. The CEFR in Practice, vol. 4. Cambridge University Press. www.cambridge.org/gb/cambridgeenglish/catalog/teacher-training-development-and-research/cefr-in-practice.
North, B. (2014b). Putting the common European framework of reference to good use. Language Teaching 47 228–249. https://doi.org/10.1017/S0261444811000206.
North, B. and Piccardo, E. (2018). Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of Reference (CEFR). Research report. Ottawa, ON: Centre for Canadian Language Benchmarks.
OECD . (2016). PISA 2015 Results (Volume I): Excellence and Equity in Education. Paris, France: OECD Publishing.
OECD . (2019). The Survey of Adult Skills: Reader's Companion, Third Edition. Paris, France: OECD Publishing.
O'Loughlin, K. (2008). The use of IELTS for university selection in Australia. IELTS Research Reports 8: 145–241.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. Language Testing 30 363–380. https://doi.org/10.1177/0265532213480336.
Pearson (2019). Score Comparison vs Other Tests for Researchers. PTE Academic. Retrieved from https://pearsonpte.com/organizations/researchers/score-comparison-vs-competitors/.
Pearson, W. S. (2019). Critical perspectives on the IELTS test. ELT Journal 73 197–206. https://doi.org/10.1093/elt/ccz006.
Peirce, B. N. and Stewart, G. (1997). The development of the Canadian language benchmarks assessment. TESL Canada Journal/ La Revue TESL Du Canada 14: 17–31.
Princeton University . (2019). International Students. Princeton University Admission. Retrieved from https://admission.princeton.edu/how-apply/international-students.
Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). Papers in Language Testing and Assessment 2: 1–27.
Ricardo-Osorio, J. G. (2008). A study of Foreign language learning outcomes assessment in U.S. Undergraduate education. Foreign Language Annals 41 590–610. https://doi.org/10.1111/j.1944-9720.2008.tb03319.x.
Rocca, L. , Carlsen, C. H. and Deygers, B. (2019). Linguistic Integration of Adult Migrants: Requirements and Learning Opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants. Council of Europe.
Sarich, E. (2012). Accountability and external testing agencies. Language Testing in Asia 2: 26. https://doi.org/10.1186/2229-0443-2-1-26.
Singer, J. D. , Braun, H. I. and Chudowsky, N. (eds.). (2018). International Education Assessments. Cautions, Conundrums, and Common Sense. Washington, DC: National Academy of Education.
Sjøberg, S. (2015). PISA and global educational governance – A critique of the project, its uses and implications. Eurasia Journal of Mathematics, Science and Technology Education 11 111–127. https://doi.org/10.12973/eurasia.2015.1310a.
Sollenberger, H. E. (1978). Development and current use of the FSI Oral Interview Test. In J. L. D. Clark (ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ/: Educational Testing Service, 1–13.
Spolsky, B. (1995). Measured Words: The Development of Objective Language Testing. Oxford, UK: Oxford University Press.
Stanford University . (2019). Exam Requirements for International Applicants. Stanford Graduate Admissions. Retrieved from https://gradadmissions.stanford.edu/applying/starting-your-application/required-exams/exam-requirements-international-applicants.
Stein, Z. (2016). Social Justice and Educational Measurement. London and New York: Routledge.
Takala, S. , Erickson, G. and Figueras, N. (2013). International assessments. In The Companion to Language Assessment. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla052/abstract.
Taylor, L. (2004). IELTS, Cambridge ESOL examinations and the common European framework. Cambridge English: Research Notes 18: 2–3.
Trim, J. L. M. (2012). The common European framework of reference for languages and its background: A case study of cultural politics and educational influences. In M. Byram and L. Parmenter (eds.), The Common European Framework of Reference: The Globalisation of Language Education Policy. Buffalo, NY/: Multilingual Matters, 14–36.
UNESCO . (2018). Global Education Monitoring Report 2019: Migration, Displacement and Education – Building Bridges, Not Walls. Geneva, Switzerland: UNESCO.
University of California . (2019). English Language Proficiency (TOEFL/IELTS). UC Admissions. Retrieved from https://admission.universityofcalifornia.edu/admission-requirements/international-applicants/english-language-proficiency-toefl-ielts.html.
University of Cambridge, L. (2017). English Language Requirements [Text]. Retrieved from www.undergraduate.study.cam.ac.uk/international-students/english-language-requirements.
University of Chicago . (2019). English Proficiency Testing. College Admissions. Retrieved from http://collegeadmissions.uchicago.edu/apply/first-year-applicants/international-applicants/english-proficiency-testing.
University College London . (2018). English Language Entry Requirements. International Students. Retrieved from www.ucl.ac.uk/prospective-students/international/applying-ucl/english-language-entry-requirements.
University of Oxford . (2019). English Language Requirements. University of Oxford. Retrieved from www.ox.ac.uk/admissions/undergraduate/applying-to-oxford/for-international-students/english-language-requirements?wssl=1.
Van Ek, J. A. (1975). Systems Development in Adult Language Learning: The Threshold Level in a European-Unit/Credit System for Modern Language Learning by Adults. Strasbourg, France: Council of

Europe.
van Ek, J. A. and Trim, J. L. M. (1991a). Threshold Level 1990. Strasbourg, France: Council of Europe.
van Ek, J. A. and Trim, J. L. M. (1991b). Waystage 1990. Strasbourg, France: Council of Europe.
van Ek, J. A. and Trim, J. L. M. (2001). Vantage. Cambridge, UK: Cambridge University Press.
Weigle, S. C. and Malone, M. M. (2016). Assessment of English for academic purposes. In K. Hyland and P. Shaw (eds.), The Routledge Handbook of English for Academic Purposes., 165–177. Routledge Retrieved from www.book2look.com/book/95iF77Y6oe.
Weir, C. J. (2003). A survey of the history of the certificate of proficiency in English (CPE) in the twentieth century. In C. J. Weir and M. Milanovic (eds.), Continuity and Innovation: The History of the CPE 1913–2002. Cambridge, UK/: Cambridge University Press, 1–56.
Wu, R. (2014). Validating Second Language Reading Examinations: Establishing the Validity of the GEPT Through Alignment with the Common European Framework of Reference, vol. 41. Cambridge, UK: Cambridge University Press.
Yale College . (2019). Applying to Yale as an International Student. Yale College Undergraduate Admissions. Retrieved from https://admissions.yale.edu/applying-yale-international-student.
Zhao, W. , Wang, B. , Coniam, D. and Xie, B. (2017). Calibrating the CEFR against the China standards of English for college English vocabulary education in China. Language Testing in Asia 7: 5. https://doi.org/10.1186/s40468-017-0036-1.
Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. Journal of Educational Change 21 245–266. https://doi.org/10.1007/s10833-019-09367-x.

Quality management in test production and administration

Wild, C. L. and Ramaswamy, R. (2008). Improving Testing. Applying Process Tools and Techniques to Assure Quality. London, UK: Routledge. This is an edited volume of 18 chapters focusing on a range of process tools and techniques to assure quality and to improve testing. It is wide ranging in coverage, and, although not specifically aimed at language testing, it deals with many of the issues raised in this chapter in an accessible way. The editors themselves set the scene by discussing the risks and the costs of poor quality in testing, and they round off the volume with thoughts on the future of quality in the testing industry.
AERA, APA and NCME . (2014). Standards for Educational and Psychological Testing. Washington, DC: AERA Publishing.
Alderson, J. C. , Clapham, C. and Wall, D. (1995). Language Test Construction and Evaluation. Cambridge, UK: Cambridge University Press.
ALTE . (1994). ALTE Code of Practice. Retrieved from www.alte.org/resources/Documents/code_practice_en.pdf.
Bachman, L. F. and Palmer, A. (1996). Language Testing in Practice. Oxford, UK: Oxford University Press.
Bachman, L. F. and Palmer, A. (2010). Language Assessment in Practice. Oxford, UK: Oxford University Press.
British Standards Institute (BSI) . (2012). Pocket Guide to Standards Development. London, UK: British Standards Institute.
Brown, J. D. (2016). Language testing and technology. In F. Farr and L. Murray (eds.), The Routledge Handbook of Language Learning and Technology. Abingdon, UK/: Routledge, 141–159.
Carr, N. T. (2014). Computer-automated scoring of written responses. In A. J. Kunnan (ed.), The companion to language assessment. Vol. 2, 1063–1078. Chichester, UK: Wiley.
Deming, W. E. (1986). Out of the Crisis. Cambridge, UK: Cambridge University Press.
Downing, S. M. and Haladyna, T. M. (2006). Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum.
Foster, D. , Maynes, D. and Hunt, B. (2008). Using data forensic methods to detect cheating. In C. L. Wild and R. Ramaswamy (eds.), Improving Testing. Applying Process Tools and Techniques to Assure Quality. London, UK: Routledge.
Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment – An Advanced Resource Book. Abingdon, UK: Routledge.
Fulcher, G. and Davidson, F. (2009). Test architecture, test retrofit. Language Testing 26 123–144. https://doi.org/10.1177/0265532208097339.
Hatch, M. J. and Cunliffe, A. L. (2006). Organization Theory: Modern, Symbolic and Post-Modern Perspectives, 2nd edn. Oxford, UK: Oxford University Press.
Heyworth, F. (2013). Applications of quality management in language education. Language Teaching 46 281–315. https://doi.org/10.1017/S0261444813000025.

Hughes, A. (2003). Testing for Language Teachers, 2nd edn. Cambridge, UK: Cambridge University Press.
ISO (International Organization for Standardization) (2015). Quality Management Systems - Requirements (ISO standard no. 9001:2015). Retrieved July 18, 2021, from https://www.iso.org/standard/62085.html
ISO (International Organization for Standardization) (2019). ISO 9001:2015 How to use it. Geneva, Switzerland: ISO. Retrieved July 18, 2021, from https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100373.pdf
Joint Committee on Testing Practices . (1988/2005). Code of fair testing practices in education (revised). Educational Measurement: Issues and Practice 24 23–26. https://doi.org/10.1111/j.1745-3992.2005.00004.x.
Jones, E. (forthcoming). Technology and cheating on tests. In G. Yu and J. Xu (eds.), Language Test Validation in a Digital Age. Studies in Language Testing Volume 52. Cambridge, UK: UCLES/Cambridge University Press.
Kane, M. (2006). Validation. In R. L. Brennan (ed.), Educational Measurement, 4th edn. Washington, DC: American Council on Education/Praeger, 17–64..
Kemp, S. (2006). Quality Management Demystified. New York, NY: McGraw Hill.
Kuijper, H. (2003). QMS as a continuous process of self-evaluation and quality improvement for testing bodies. Retrieved August 2, 2011, from www.alte.org/qa/index.php.
Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (ed.), Fairness and Validation in Language Assessment. Cambridge, UK/: Cambridge University Press, 1–13.
Kunnan, A. J. (2004). Test fairness. In M. Milanovic and C. Weir (eds.), European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001. Cambridge, UK/: UCLES/Cambridge University Press, 27–50.
Lane, S. , Raymond, M. R. and Haladyna, T. M. (2016). Handbook of Test Development, 2nd edn. Abingdon, UK: Routledge.
Lane, S. , Raymond, M. R. , Haladyna, T. M. and Downing, S. M. (2016). Test development process. In S. Lane , M. R. Raymond and T. M. Haladyna (eds.), Handbook of Test Development, 2nd edn. Abingdon, UK: Routledge, 3–18.
Link, S. (forthcoming). The interface between automated essay scoring and socio-cognitive research. In G. Yu and J. Xu (eds.), Language Test Validation in a Digital Age. Studies in Language Testing Volume 52. Cambridge, UK: UCLES/Cambridge University Press.
Mayer-Schönberger, V. and Cuckier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. Boston, MA: Houghton Mifflin Harcourt.
McNeely, C. L. and Hahm, J. (2014). The big (data) bang: Policy, prospects and challenges. Review of Policy Research 31 304–310. https://doi.org/10.1111/ropr.12082.
Messick, S. (1989). Validity. In R. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Macmillan, 13–103.
Nakatsuhara, F. , Inoue, C. , Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. Language Assessment Quarterly 14: 1–18.
National Research Council (NRC) . (2013). Frontiers in Massive Data Analysis. Washington, DC: The National Academies Press.
O'Reilly. (2011). Velocity 2011: Jon Jenkins, "Velocity Culture" [Video File]. Retrieved from www.youtube.com/watch?v=dxk8b9rSKOo.
Richards, N. M. and King, J. H. (2014). Big data ethics. Wake Forest Law Review 49: 393–432.
Saville, N. (2005). Setting and monitoring professional standards: A QMS approach. Research Notes 22 2–5. Cambridge, UK: Cambridge ESOL.
Saville, N. (2010). Auditing the quality profile: From code of practice to standards. Research Notes 39 24–28. Cambridge, UK: Cambridge ESOL.
Saville, N. (2014). Using Standards and Guidelines. In A. J. Kunnan (ed.), The Companion to Language Assessment: Approaches and Development Volume II. Oxford, UK: Wiley Blackwell.
Saville, N. (2016). Principles of Good Practice. Cambridge, UK: Cambridge English Language Assessment.
Saville, N. and Hargreaves, P. (1999). Assessing speaking in the revised FCE. ELT Journal 53 42–51. https://doi.org/10.1093/elt/53.1.42.
Shewhart, W. A. (1931). Economic Control of Quality of Manufactured Product. New York, NY: D. Van Nostrand Company.
Shewhart, W. A. (1939). Statistical Method from the Viewpoint of Quality Control. Washington, DC: The Graduate School, the Department of Agriculture.
Taylor, F. W. (1911). The Principles of Scientific Management. New York, NY: Harper and Brothers Publishers.
van Avermaet, P. (2003). QMS and The Setting of Minimum Standards: Issues of Contextualisation Variation Between the Testing Bodies. Retrieved August 2, 2011, from www.alte.org.

van Avermaet, P. , Kuijper, H. and Saville, N. (2004). A code of practice and quality management system for international language examinations. Language Assessment Quarterly 1 137–150. https://doi.org/10.1080/15434303.2004.9671781.
Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke, UK: Palgrave Macmillan.
Wild, C. L. and Ramaswamy, R. (2008). Improving Testing: Applying Process Tools and Techniques to Assure Quality. London, UK: Routledge.
Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Retrieved from www.nature.com/articles/sdata201618.
Yu, G. and Xu, J. (eds.). (forthcoming). Language Test Validation in a Digital Age. Studies in Language Testing Volume 52. Cambridge, UK: UCLES/Cambridge University Press.

Epilogue

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. Language Testing. https://doi.org/10.1177/0265532220951504.
Bennett, R. E. (2000). Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing. Princeton, NJ: ETS.
Deygers, B. and Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. Language Testing 36 347–368. https://doi.org/10.1177/0265532219826390.
Eberharter, K. (2021). An Investigation into the Rater Cognition of Novice Raters and the Impact of Cognitive Attributes When Assessing Speaking. Unpublished doctoral dissertation, Lancaster University, United Kingdom.
Fulcher, G. (2012). Assessment literacy for the language classroom. Language Assessment Quarterly 9 113–132. https://doi.org/10.1080/15434303.2011.642041.
Fulcher, G. (2015). Re-Examining Language Testing: A Philosophical and Social Inquiry. Oxon, UK: Routledge.
Fulcher, G. (2020). Operationalizing assessment literacy. In D. Tsagari (ed.), Language Assessment Literacy: From Theory to Practice. Newcastle upon Tyne, UK/: Cambridge Scholars, 8–28.
Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York: Routledge.
Gebril, A. (ed.). (2021). Learning-Oriented Language Assessment: Putting Theory into Practice. New York, NY: Routledge.
Green, A. and Van Moere, A. (2020). Repeated test-taking and longitudinal test score analysis. Language Testing 37 475–481. https://doi.org/10.1177/0265532220934202.
Gruba, P. (2020). What does language testing have to offer to multimodal listening? In G. J. Ockey and B. A. Green (eds.), Another Generation of Fundamental Considerations in Language Assessment. Singapore/: Springer, 43–57.
Harding, L. and Kremmel, B. (forthcoming). Technology, values, ethics and consequences: From innovation to impact in language assessment.
Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang , R. I. Thomson and J. Murphy (eds.), The Routledge Handbook of Contemporary English Pronunciation. London, UK/: Routledge, 570–584.
Isbell, D. and Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high stakes decisions. Language Testing 37 600–619. https://doi.org/10.1177/0265532220943483.
Knoch, U. , Huisman, A. , Elder, C. , Kong, X. and McKenna, A. (2020). Drawing on repeat test takers to study test preparation practices and their links to score gains. Language Testing 37 550–572. https://doi.org/10.1177/0265532220927407.
Kremmel, B. and Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. Language Assessment Quarterly 17 100–120. https://doi.org/10.1080/15434303.2019.1674855.
Lamprianou, I. , Tsagari, D. and Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. Language Testing. https://doi.org/10.1177/0265532220940960.
Marra, A. , Buonanno, P. , Vargas, M. et al. (2020). How COVID-19 pandemic changed our communication with families: Losing nonverbal cues. Critical Care 24: 297. https://doi.org/10.1186/s13054-020-03035-w.
McNamara, T. and Roever, C. (2006). Language Testing: The Social Dimension. Malden, MA and Oxford, UK: Blackwell Publishing Ltd.

Meurers, D. , De Kuthy, K. , Nuxoll, F. , Rudzewitz, B. and Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. Annual Review of Applied Linguistics 39 161–188. https://doi.org/10.1017/S0267190519000126.

Mirhosseini, S. A. and De Costa, P. (eds.). (2020). The Sociopolitics of English Language Testing. London, UK and New York, NY: Bloomsbury.

Nakatsuhara, F. , Inoue, C. , Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. Language Assessment Quarterly 14 1–18. https://doi.org/10.1080/15434303.2016.1263637.

Ockey, G. J. , Gu, L. and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. Language Assessment Quarterly 14 346–359. https://doi.org/10.1080/15434303.2017.1400036.

Pill, J. and Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. Language Testing 30 381–402. https://doi.org/10.1177/0265532213480337.

Shohamy, E. G. (2001). The Power of Tests: A Critical Perspective on the Uses of Language Tests. London: Longman/Pearson Education.

Sultana, N. (2019). Exploring the Alignment of the Secondary School Certificate English Examination with Curriculum and Classroom Instruction: A Washback Study in Bangladesh. Unpublished doctoral dissertation, Queen's University, Kingston, Ontario, Canada). Retrieved from https://qspace.library.queensu.ca/handle/1974/26482.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. Language Testing 30 403–412. https://doi.org/10.1177/0265532213480338.

Taylor, L. and Harding, L. (2020). A testing time for testing: Assessment literacy as a force for social good in the time of coronavirus. Campaign for Social Science. Retrieved January 19, 2021, from https://campaignforsocialscience.org.uk/news/a-testing-time-for-testing-assessment-literacy-as-a-force-for-social-good-in-the-time-of-coronavirus/.

, Turner C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (eds.), Handbook of Second Language Assessment. Boston and Berlin/: De Gruyter, Inc, 255–272.

Yan, X. and Fan, J. (2021). "Am I qualified to be a language tester?": Understanding the development of language assessment literacy across three stakeholder groups. Language Testing 38. https://doi.org/10.1177/0265532220929924.