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*. 

The Routledge Handbook of 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
## Contents

- List of figures ix
- List of tables x
- List of contributors xi
- Acknowledgements xviii

### Editorial

Glenn Fulcher and Luke Harding

### SECTION 1

#### Validity

1  Conceptions of validity 17
   Carol A. Chapelle and Hye-won Lee

2  Articulating a validity argument 32
   Michael T. Kane

3  Inference and prediction in language testing 48
   Steven J. Ross

### SECTION 2

#### The uses of language testing

4  Social dimensions of language testing 63
   Richard F. Young

5  Designing language tests for specific purposes 81
   Carol Lynn Moder and Gene B. Halleck

6  Revisiting language assessment for immigration and citizenship:
   The case of US citizenship and the Naturalization Test 96
   Antony John Kunnan
SECTION 3
Classroom assessment and washback 117

7 Classroom-based assessment 119
   Janna Fox, Nwara Abdulhamid, and Carolyn E. Turner

8 Washback: Looking backward and forward 136
   Liying Cheng and Nasreen Sultana

9 Assessing young learners 153
   Yuko Goto Butler

10 Dynamic assessment 171
   Marta Antón and Próspero N. García

11 Diagnostic assessment in language classrooms 187
   Eunice Eunhee Jang and Jeannie Sinclair

SECTION 4
Assessing the language skills 207

12 Assessing speaking 209
   Fumiyo Nakatsuhara, Nahal Khabbazbashi, and Chihiro Inoue

13 Assessing listening 223
   Elvis Wagner

14 Assessing writing 236
   Ute Knoch

15 Assessing reading 254
   Tineke Brunfaut

SECTION 5
Test design and administration 269

16 Test specifications 271
   Yan Jin

17 Evidence-centered design in language testing 289
   Chengbin Yin and Robert J. Mislevy
18 Accommodations and universal design
   Jamal Abedi

19 Rater and interlocutor training
   Larry Davis

SECTION 6
Writing items and tasks

20 Item writing and item writers
   Dongil Shin

21 Writing integrated tasks
   Lia Plakans

22 Test-taking strategies and task design
   Andrew D. Cohen

SECTION 7
Prototyping and field tests

23 Prototyping new item types
   Susan Nissan and Elizabeth Park

24 Pre-operational testing
   Benjamin Kremmel, Kathrin Eberharter, and Franz Holzknecht

25 Piloting vocabulary tests
   John Read

SECTION 8
Measurement theory in language testing

26 Classical test theory
   James Dean Brown

27 Item response theory and many-facet Rasch measurement
   Gary J. Ocket

28 Reliability and dependability
   Xun Yan and Jason Fan
# Contents

29  Scoring performance tests  
    *Evelina D. Galaczi and Gad S. Lim*  
    495  

## SECTION 9  
Technology in language testing  

30  Validity and the automated scoring of performance tests  
    *Xiaoming Xi*  
    513  

31  Computer-based testing  
    *Yasuyo Sawaki*  
    530  

32  Corpus linguistics and language testing  
    *Sara T. Cushing*  
    545  

## SECTION 10  
Ethics, fairness, and policy  

33  Ethics and fairness  
    *F. Scott Walters*  
    563  

34  Standards in language proficiency measurement  
    *Bart Deygers*  
    578  

35  Quality management in test production and administration  
    *Nick Saville and Sarah McElwee*  
    597  

36  Epilogue: language testing: where are we heading?  
    *Luke Harding and Glenn Fulcher*  
    622  

*Index*  

633
Figures

3.1 Scatterplot of grade-point average with truncated aptitude 54
3.2 Scatterplot of grade-point average with low-aptitude trainees 55
3.3 Parallel growth model for proficiency and aptitude 56
6.1 An example of the literacy test in Italian 103
10.1 Sample CDA learner profile 179
14.1 Process of writing assessment 237
14.2 Levels of writing development 247
16.1 A comparison of specification-driven and effect-driving testing 283
17.1 Toulmin’s general structure for arguments 290
17.2 Extended Toulmin diagram for assessment 292
23.1 Schematic table 404
23.2 Summary task 405
25.1 Sample item from the WAT (translated from Dutch) 438
27.1 ICCs for three items from the 1PL Rasch model 466
27.2 ICCs for 1PL, 2PL, and 3PL 469
27.3 Category response curves for writing example 471
28.1 Some sources of error in a test 483
28.2 Two sample multiple-choice items measuring vocabulary 484
28.3 Scatterplots of test scores on two administrations 485
32.1 Output of CLAWS tagger 548
32.2 Output of USAS tagger 549
32.3 Output from Stanford parser 550
32.4 Output of concordance on truth 551
33.1 Worksheet grid 574
35.1 Core processes and outputs of the assessment cycle 606
35.2 The assessment cycle 608
35.3 The assessment cycle showing periodic test review 613
### Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Characteristics defining validity as presented by Messick (1989)</td>
<td>19</td>
</tr>
<tr>
<td>1.2</td>
<td>Types of evidence pertaining to test score interpretations and actions from Messick (1989) and in the Standards (1999, 2014)</td>
<td>20</td>
</tr>
<tr>
<td>1.3</td>
<td>Types of claims in validity arguments and the inferences leading to each type</td>
<td>22</td>
</tr>
<tr>
<td>6.1</td>
<td>Applications, naturalizations, and denials, 2015–2019</td>
<td>109</td>
</tr>
<tr>
<td>6.2</td>
<td>Naturalizations from six leading countries by year and percentage, 2015–2019</td>
<td>109</td>
</tr>
<tr>
<td>10.1</td>
<td>Pre-scripted prompts for classroom DA</td>
<td>177</td>
</tr>
<tr>
<td>10.2</td>
<td>Functions of the scores and learner profile automatically generated by the C-DA tests</td>
<td>179</td>
</tr>
<tr>
<td>14.1</td>
<td>Dimensions and considerations in writing assessment</td>
<td>238</td>
</tr>
<tr>
<td>14.2</td>
<td>Continuum of writing constructs in different assessment types</td>
<td>245</td>
</tr>
<tr>
<td>14.3</td>
<td>Continuum of level of detail in scoring</td>
<td>245</td>
</tr>
<tr>
<td>14.4</td>
<td>Continuum of score and feedback provision related to different assessment types</td>
<td>245</td>
</tr>
<tr>
<td>16.1</td>
<td>The SAT essay domain</td>
<td>275</td>
</tr>
<tr>
<td>16.2</td>
<td>A summary of the rubric of the SAT Essay Test</td>
<td>276</td>
</tr>
<tr>
<td>16.3</td>
<td>Specification of some features of the CET-SET Band 4</td>
<td>276</td>
</tr>
<tr>
<td>17.1</td>
<td>Summary of ECD layers in the context of language testing</td>
<td>291</td>
</tr>
<tr>
<td>17.2</td>
<td>Design pattern attributes and relationships to assessment argument</td>
<td>293</td>
</tr>
<tr>
<td>17.3</td>
<td>A design pattern for assessing cause-and-effect reasoning reading comprehension</td>
<td>294</td>
</tr>
<tr>
<td>17.4</td>
<td>Steps taken to redesign TOEFL iBT and TOEIC speaking and writing tests and guided by layers in evidence-centered design</td>
<td>299</td>
</tr>
<tr>
<td>27.1</td>
<td>Test taker response on a multiple-choice listening test</td>
<td>464</td>
</tr>
<tr>
<td>28.1</td>
<td>Classification of test takers on two test administrations</td>
<td>487</td>
</tr>
<tr>
<td>28.2</td>
<td>An example of classification of test takers for computing kappa coefficient</td>
<td>488</td>
</tr>
<tr>
<td>34.1</td>
<td>PISA reading performance levels</td>
<td>580</td>
</tr>
<tr>
<td>34.2</td>
<td>Kaulfers’s aural performance scale</td>
<td>582</td>
</tr>
<tr>
<td>34.3</td>
<td>FSI/ILR speaking descriptors (1968)</td>
<td>583</td>
</tr>
<tr>
<td>34.4</td>
<td>ACTFL alignment with ILR scales and CEFR levels</td>
<td>584</td>
</tr>
<tr>
<td>34.5</td>
<td>Organization of CEFR levels</td>
<td>585</td>
</tr>
<tr>
<td>34.6</td>
<td>Overview of the CLB and their relationship with the CEFR levels</td>
<td>587</td>
</tr>
<tr>
<td>34.7</td>
<td>CSE overview and CEFR link</td>
<td>588</td>
</tr>
<tr>
<td>34.8</td>
<td>Minimum English language proficiency requirements for international students in arts and humanities</td>
<td>589</td>
</tr>
<tr>
<td>34.9</td>
<td>Score equivalence guides compared</td>
<td>589</td>
</tr>
</tbody>
</table>
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*
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.
Contributors

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.
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.
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.
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 rea-
son, 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 qual-
ity 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 there-
fore 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 chap-
ter 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 statisti-
cal 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 com-
monplace 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 devel-
opment 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

Conceptions of validity


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.


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.


Inference and prediction in language testing


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.


Designing language tests for specific purposes


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.


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.


Revisiting language assessment for immigration and citizenship
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.
that invents “persons and communities with respect to space and place” (p. 1).
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.
Classroom-based assessment


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).


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.


Assessing young learners


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.


University Press, 70–127.


Dynamic assessment


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


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.


---

**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.


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.


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.


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.


Test specifications


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.


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.


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-cognitive systems perspective on language.


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.


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 an 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 ELLs' language barriers, it provides content definitions as well, which is unfair to the recipients of this accommodation. Based on the review of literature, we identified five different language-based accommodations that do not provide an 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.

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 rigor 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.


Exceptional Children.


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.


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. Shohamy’s claims that intended effects can lead to unintended and unexpected consequences deserve thoughtful attention in all areas of test design.


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., Englehard, G., Raczyński, 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.


**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).


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.


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 particular note.


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.


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-operative testing issues can be situated in a validation argument.


**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.


Classical test theory


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.


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.
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.
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.


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.


Scoring performance tests
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 speaking assessment and provides an overview of recent developments and future outlook.


**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 Language™. 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.


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.


Barlow, M. (2012). MonoConc Pro 2.2 (MP2.2) [Software].


**Ethics and fairness**


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.


Standards in language proficiency measurement


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.


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.


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.


Epilogue


