John Pill

Book review

Introducing Second Language Assessment
by Gary J. Ockey. Cambridge: Cambridge University Press, 2025, 266 pp., GBP27.99 (digital), ISBN 9781009067461. https://doi.org/10.1017/9781009067461

This textbook introducing second language assessment is part of the series Cambridge Introductions to Applied Linguistics. The target readers are "graduate and advanced undergraduate students in language teacher training courses, students in TESOL/TEFL certificate programs, and practicing second language teachers" (p. xxi), and the text meets the needs of these groups through its well-structured content and an explicit focus on raising awareness of language assessment literacy. The textbook presents examples mainly from English-language assessments, so making them accessible to all readers, but the theory and methods introduced will apply to assessment of any language. It starts with an overview of the types and purposes of language assessment in educational settings and then focuses on the practical selection and creation of test tasks. I find the book offers a strong foundation for newcomers to language assessment. In terms of content, the book's focus is tighter than similar recent textbooks on practical language assessment (e.g., Brown & Abeywickrama, 2018; Fulcher, 2025; Green, 2021; Hughes & Hughes, 2020). It shares with these texts the goals of helping language teachers make good test tasks and providing the reassurance of a solid framework to support this endeavour. It offers a distillation of language testing knowledge and practice as established over the last three decades and adds consideration of current technologies such as generative artificial intelligence (AI), a topic covered only in a limited way in other introductory publications.

The main textbook has five parts and 13 chapters. Their sequence creates a coherent structure for the text as a whole. Part I establishes the context and motivation for the practice of language assessment. Chapter 1 provides examples of how language assessments have been used (and often misused) in different contexts, and it prompts readers to consider what they need to know about the theory and practice of language assessment in their current or future role as language teachers, which is termed their language assessment literacy. Chapter 2 sets out different types of language assessment, stressing the importance of matching a tool with its purpose. Ockey's two overall aims for language assessment are presented: "providing an accurate indication of a test taker's language ability or achievement and promoting effective language learning practices" (p. 17). Readers are reminded of these aims several times during the text.

Part II of the book presents broad language assessment principles. Chapter 3 considers factors that affect where and when tests are used, for what purpose, and whether they succeed. Chapter 4 introduces concepts of validity and alignment and includes a useful section on the ways constructs, the abilities that language assessments seek to measure, might be defined. Chapter 5 deals with the goal of assessment consistency, distinguishing uniformity (aspects of design and delivery) from reliability (scoring). Having set the foundations in Part II, the text then turns to more practical topics.

The focus for Part III is designing or selecting dichotomously scored items and tasks to measure comprehension (reading and listening skills) and knowledge (grammar and vocabulary) and how to evaluate test-taker responses to them. Chapter 6 presents different item types and foregrounds the need for suitable input texts for tests of the receptive skills.

Chapter 7 demonstrates how content analysis and descriptive statistics are used to evaluate the appropriateness and effectiveness of norm-referenced tests (NRT; tests that compare and rank test takers). Then, in Chapter 8, content and statistical analysis are used to determine quality at the item level. The statistics Item Facility and Point-Biserial are introduced and illustrated using test data, so that statistical results can be interpreted by linking them to the items that produced them. Reliability, relating to the internal consistency of NRTs, is also introduced in Chapter 8, using Cronbach's Alpha. Chapter 9 attends to parallel tools and interpretations for criterion-referenced tests (CRT; tests that measure test takers against pre-determined standards of skill or knowledge). Ways of standard-setting to determine cut-scores are presented, and CRT-relevant statistics and concepts are explained (e.g., B-Index and dependability in place of NRTs' Point-Biserial and reliability).

Part IV covers performance assessment scored polytomously and, in its format, parallels Part III's coverage of dichotomously scored items. It focuses on task types used to elicit samples of speech and writing from test takers, and how these language performances are to be assessed appropriately and consistently. In Chapter 10, a range of task types is set out, and then consideration is given to how performance assessments are delivered, in person and online. Chapter 11 revisits scoring, here describing holistic and analytic rating schemes and how these may be developed and then applied by human raters, automated systems, or both. Chapter 12 covers how to judge the effectiveness – reliability and dependability – of human-scored performance assessments, considering situations in which tests can be considered both norm- and criterion-referenced.

Part V is titled Reflecting and Self-Assessment and contains the final chapter, in which readers are encouraged to re-examine their own language assessment literacy having worked through the book's content. This review includes a Language Assessment Literacy Test, comprising test tasks with prompts to guide a critique of them. The textbook concludes with a useful glossary of terms related to language assessment, a list of references and an index.

Several pedagogically useful features are included, making the book attractive for instructors working with teachers in training (the target audience). Regular appearance in each chapter increases their value and contributes to the clear structuring of content. Within the main text, features include: *Time to think* boxes, with questions and prompts encouraging readers to relate topics to their own experience; sets of discussion points and exercises at the end of each chapter; recommendations of additional resources to extend the topic, mainly through academic articles and online videos involving language testing experts; and, for several of the practice-focused chapters, appendices with explanations of how to use computer software (Excel) to carry out statistical calculations introduced in the text, and generative AI technologies (ChatGPT) to help develop input texts, items, prompts and rating scales. A further feature of the book that is likely to support effective learning are the sets of guidelines that synthesize key points and so provide practical tools to evaluate test materials. External to the text, the publisher's website also offers resources for students and instructors. Students can access spreadsheets (Excel) and video instructions for the statistical calculations in the book; video guides are also provided for using ChatGPT in test creation. Instructor resources (with restricted access) include a teacher's guide with answers to the exercises in the main book, and slide decks (PowerPoint) presenting core content of each chapter.

The description above indicates that the structure of the textbook is strong. Furthermore, its content is chosen with great care and is presented systematically. The methodical approach to presenting content means that all aspects relevant to selecting or designing a test are included,

including areas that might be less commonly considered, for example (as already mentioned), statistics and concepts to support the use of criterion-referenced tests in Chapter 9. Content is presented using worked examples of tests – input texts, tasks, and test takers' responses and performances – that are skilfully selected across the chapters to illustrate a range of test types (for several skills; independent, integrated and interactive tasks) and test uses in education settings. I note that a similar level of attention is given to areas about which it is often difficult to be definitive, for example, ways to define a construct or create a rating scale systematically. In this text, the author presents several clear and practical options to address each of these topics (e.g., four options for rating scales on p. 187).

A further notable structural element is the anecdote at the start of each chapter, drawn from Ockey's long experience as a language educator in the United States and elsewhere. Each anecdote provides an entry into the topic to be covered and connects with the reality of language assessment, recognising that situations are often less than ideal and mistakes can be made. It is then satisfying how, as it concludes, each chapter returns to this story with a resolution or further commentary. This looping back also occurs at a higher level, with readers being asked to measure their own language assessment literacy at the start and end of the book, using a version of Kremmel and Harding's (2020) framework of language assessment literacy. It would be interesting to try this with a class of language teachers in training over a semester, although, as Ockey notes (p. 219), readers may overestimate their knowledge at the start. (They may also stumble over the scale to record their self-evaluation, as using percentages to indicate what I know or need to know does not seem an intuitive approach.) These features of the text design indicate that the book is better studied from beginning to end, certainly on first reading, than dipped into selectively. Readers should trust the author's sequence of presentation in the practice-focused chapters (the "staging" of the content) – an element that does not seem quite right at the start of a chapter (e.g., an example with a poorly worded multiple-choice option) is unlikely to be present by accident, and it will be dealt with by its end. The evident care taken by the author in planning the content should be summarised for readers in a map of the book at the start, indicating how the various threads are woven together. The linear presentation in the conventional contents page cannot capture the clever underlying design which may itself guide newcomers to understand the field.

The text considers topics of practical and current interest in test development. I welcome how the examples given of test task evaluation based on responses from test takers foreground the need to consider content analysis of the texts/items and statistical analysis together. Readers are walked through the experience of noticing an unusual statistic and finding out what caused it. As another example, there is discussion of the authenticity of input texts for listening tasks and the use of different speech varieties in these recordings, with readers asked to consider how a test should represent its target language use situation. There is also engagement with developments in technology. Aspects of delivering tests online are covered throughout, and the use of generative AI in test development is considered in practical terms (as noted already). However, the use of automated scoring is covered only partially. As an instance of where coverage could be expanded, it is not explained how computer scoring in a writing assessment for Saudi young learners of English would deal with the criterion of "handwriting neatness" (pp. 192-193) when the test takers' scripts are written on paper in the examples provided. It would be possible to extend this example to illustrate how human raters and automated scoring systems work in very different ways to provide scores assumed to be comparable. Likewise, the guidelines on creating or evaluating performance assessment

tasks include “The scoring criteria are clear” (p. 203), but the opportunity is missed to open up discussion of how this quality might be evaluated when automated scoring is used.

Although the scope of the book is clear, allowing good coverage of what it sets out to deliver, a broader view might be worthwhile. For example, the focus on selecting and developing tests at the task level may be limiting for readers, who may also need guidance on how to combine tasks and tests focused on one skill to create more comprehensive tests of several aspects of language proficiency for use in their classrooms. Similarly, readers are likely to work in contexts where new tests are required regularly, perhaps once every semester or year, so advice also seems appropriate on procedures to sustain the development of a series of test forms – a stronger sense of a *test design cycle* (Green & Fulcher, 2021). While the book covers the topics of piloting test tasks and learning from analysis of test performance to inform future tests, the ongoing nature of this work is not emphasised. Other topics might also be given greater prominence. For instance, the benefits are not stressed of teacher collaboration to develop assessments, although teamwork is illustrated in the anecdotes and assumed in the exercises set at the end of chapters. Furthermore, greater attention might be given to the resources required for test development, including suggestions of options to consider where resources are limited. In this context, practical constraints, such as copyright restrictions and ethical issues around using recordings of authentic conversations as input texts, might also be mentioned.

Some of the principles introduced at the start of the book are not revisited explicitly during the practice-focused chapters. From its title, Reflecting and Self-Assessment, I initially expected that Part V of the book would outline how teachers can guide learners in techniques for informal assessment promoting language development. Self-assessment is introduced in Chapter 2 as one purpose of language assessment, along with dynamic assessment and learning-oriented assessment (among others), but the book does not address directly how assessment may affect classroom practice, how scores and feedback can be the starting point for further learning, or how assessments might best be designed to fulfil this purpose. These seem to be issues in language assessment that teachers in training might usefully reflect on. More broadly, although examples of misuse of language tests at the start of the book are included to acknowledge the doubts held by some teachers (and other readers), concerns about the legitimacy of assessment (e.g., its use to regulate whether dependants can join a family member who has migrated to another country) are not pursued further. I wonder if the later chapters of the text do enough to persuade doubters of the benefits of language assessment, especially when applied beyond educational settings. However, this is perhaps not the author’s goal in an introductory volume.

It is important to consider a textbook’s useability for a global readership, while appreciating the impossibility of writing one to suit all contexts. The text generally seems accessible to readers around the world, using examples from English-language assessments. Complexities that arise in practice are included – for example, assessing performances by test takers who have different interactional norms (p. 37) – but how they might be resolved is left open. Knowing the age group that a “third-grade elementary teacher” (p. 28) teaches assumes understanding of the US education system, though elsewhere in the book explanatory information is provided. The term “sections” (e.g., p. 49) for the divisions of a cohort of university students into teaching groups is also not universally understood.

A few minor typographical errors are noted here with the aim of helping readers follow the text. In Example 6.1 (p. 82) the four options are not labelled a, b, c, and d as the instruction

indicates. The hyperlinks to Guidelines 7.1 (end of p. 99, start of p. 102) should be to Guidelines 7.2. In the examples of holistic and analytic rating scales (Tables 11.1 and 11.2), the parallel proficiency bands scoring 1 are named Developing and Emerging, respectively, but this difference is not explained. On p. 189, a word is omitted: “Naila would get the *highest* total score possible”.

I would like to comment on the digital version of the text that I reviewed. The interface is generally easy to use. The search feature in the Cambridge Spiral eReader is useful and compensates for the long contents page (with hyperlinks) within the text and in the Table of Contents tool, both of which I found cumbersome to navigate. Hyperlinks are used to some good effect: the year of publication of an in-text citation links to the full citation in the References section; mentions of other chapters and examples within the text are hyperlinked to them, as you would expect. However, there is no obvious way (that I could find) to go back to where you have come from in the text, and so I generally avoided the links. Also, when terms are introduced in the text, their definitions (taken from the glossary) appear when you hold the mouse cursor over the highlighted text. However, on their first appearance, these terms are usually also defined in the text, removing the need for this feature. Moreover, elsewhere in the text, where quick access to the definition of an unfamiliar term might be useful, this feature is not available.

Overall, this publication is a thorough and well-designed introduction for use with the groups it targets. Its content is covered competently and clearly, and it is a highly useable resource for instructors, providing structure and support for effective teaching about how to deliver second language assessments primarily in educational contexts.

Acknowledgement

The author would like to thank the three anonymous reviewers for their generous feedback.

References

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.

Fulcher, G. (2025). *Practical language testing* (2nd ed.). Routledge.

Green. A. (2021). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Green, A., & Fulcher, G. (2021). Test design cycle. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 69-77). Routledge.

Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (3rd ed.) Cambridge University Press.

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy scale. *Language Assessment Quarterly, 17*(1), 100-120. https://doi.org/10.1080/15434303.2019.1674855