J. Fan. *The Development and Validation of Standards in Language Testing*. Fudan University Press. 2018. Pp. 257.

Since the beginning of the new millennium, the term "standards" has become a catchword in educational contexts, due largely to the standards-driven education reform in the U.S. and the application of the Common European Framework of Reference for Languages (Council of Europe, 2001). In these situations, "standards" is used in reference to "a level of performance required or experienced" (Davies et al., 1999: 185). The alignment of all the elements in an educational system with performance standards brings greater coherence and efficiency to the system (see Macqueen et al. this issue). There is, however, a second sense of the term, which refers to "a set of principles which can be used as a basis for evaluating what language testers do", leading to "codification in an agreed set of guidelines or Code of Practice" (ibid.). Such codification, according to Davies et al. (1999), indicates a concern to establish professional ethics. In the book The Development and Validation of Standards in Language Testing, Dr. Jason Fan, then associate professor at Fudan University, China, adopted this second sense of the term "standards". Funded by the National Office for Philosophy and Social Sciences (NOPSS) of China, Jason, as the principal researcher, conducted a three-year research project into the role of professional standards in language testing practice and the way of establishing the standards. In this 257-page monograph written in Chinese, Jason provides a comprehensive summary of the NOPSS project.

The book has nine chapters, which can be grouped into three parts, each focusing on one of the three phases of standards development and validation. Part One (Chapters 1 to 3) evaluates the frameworks for test development and validation and reviews major language testing standards, thus laying a solid foundation for the project. In Chapter 1, the concept of standards as used in the book is defined by reference to both the European perspective of codification (e.g., Boyd & Davies, 2002; Davies, 2008) and the standards-based testing practices in the US (e.g., AERA, APA & NCME, 1985, 1999, 2014). The aims and objectives of the NOPSS project are also described in the chapter, which include 1) reviewing the status quo in the development and use of professional standards, 2) analyzing the distinctive contextual features of language testing in China, 3) developing the standards for language test development and use in China, and 4) implementing and validating the standards. Chapter 2 elaborates on the key issues and procedures involved in language test development, followed by a discussion of two major frameworks for developing and validating language tests: Bachman and Palmer's (1996) test usefulness framework and Bachman and Palmer's (2010) assessment use argument (AUA). The latter is considered as a more suitable "guiding theoretical framework" for standards development. Chapter 3 provides a balanced overview of 14 leading professional standards in the field of language

testing, covering a wide range of topics: test takers' responsibilities and duties, code of ethics, code of practices, test development and administration, test uses, and test fairness. Five of the standards are examined in detail with respect to their background, objectives, structure and features. The chapter ends with an overview of three empirical studies of professional standards in the field of language testing.

Part Two (Chapters 4 to 6) identifies the needs of the field for professional standards through an investigation of current practices of test development and use in the Chinese context. Chapter 4 reports on a survey of the practices in developing and implementing post-entry placement tests of English. Questionnaire data were collected from 36 universities and follow-up interviews were conducted with teachers from five of the universities. Guided by the theoretical framework established in Chapter 3 (Table 3.1 of this book, p. 52), data were analyzed to understand the status quo of language testing practices and the challenges facing the test developers. Chapter 5 explores teachers' and students' perceptions of language testing practices in China. Data were collected through a student questionnaire survey (n=248), student focus group discussion (n=14), and one-on-one interviews with students (n=14) and teachers (n=12). Following the same theoretical framework, data were analyzed and problems were identified as areas of concern by both groups of respondents, including insufficient assessment of speaking, susceptibility to cheating, lack of transparency of information, and inadequate feedback on performance. Chapter 6 continues the exploration of teachers' and students' perceptions by means of large-scale surveys. A questionnaire was developed based on the qualitative findings reported in the previous two chapters and the Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014). Data were collected from university teachers (n=100) and students (n=381). The study presents a fuller picture of the contextual features of language test development and use in China. Test administration is perceived as the strength of testing practices, whereas test/task design seems to be an area where there is still considerable room for improvement.

Part Three (Chapters 7 and 8) delves into the development and validation of standards for language test development and use in the Chinese context. Based on the review of standards in Part One and the contextual analysis in Part Two, Chapter 7 presents a concise summary of the needs for professional standards by different stakeholder groups in the Chinese educational context: education and assessment management departments, test developers, test users (e.g., admission officers, employers), and key stakeholders (e.g., teachers, students). The main parts of the chapter, which are also the highlights of the book, are a detailed discussion of the purposes, methods and procedures of standards development and a clear presentation of the structure and content of the standards. Specifically, the tables and figures (i.e., Tables 7.1, 7.2; Figures 7.2, 7.3, 7.4) in the chapter are carefully designed and most useful for readers who are interested in language testing standards. Chapter 8 moves beyond standards

development to explore the application and validation of the standards. To showcase the usefulness of the standards, the chapter delineates the design and implementation of the Fudan English Test (FET), an English proficiency test for students at Fudan University, a top-tier university in China. The case analysis includes a step-by-step unfolding of the development of test specifications, item construction and modification, and quality control in scoring constructed-response items, as well as a meticulous presentation of the evidence for the construct validity and stakeholder perceptions of the FET. Finally, challenges facing the implementation of standards are discussed from the perspectives of practical constraints, assessment literacy of stakeholders, and enforcement mechanisms.

Chapter 9 concludes the book by summarizing the main findings of the project, pointing out the limitations, and suggesting future directions for research. It is stressed that standards need to be developed based on a deeper understanding of the educational context, particularly of the practices which characterize organizations or commercial agencies operating large-scale language testing programs. Future research also needs to include more variety in stakeholder groups so as to arrive at a more nuanced and balanced picture of the context in which language tests are developed and used.

To the best of my knowledge, there have been very few attempts within the profession to document the development of standards for improving the quality of language testing and assessment. The author, Dr. Jason Fan, now Deputy Director of the Language Testing Research Centre at the University of Melbourne, started his exploration of language testing standards during his doctoral study at Shanghai Jiao Tong University and has developed considerable interest and expertise in this area. In this book, a systematic approach has been adopted to establish a solid theoretical and practical foundation for the construction of standards that are most appropriate for a local context. An even more valuable part of the project is the application and validation of the standards that have been proposed in the project. De Jong and Zheng (2011) point out that "(D)espite the abundance of these standards and guidelines, how they are observed in practical testing developments is rarely documented; furthermore, there have observed an even sparse application of these standards or guidelines in practical test development practices" (p. 1). In a paper Jason and I co-authored (Jin & Fan, 2015), quoting Boyd and Davies (2002), we argue that language testing is a weak profession compared with the law profession or the medical profession, because ethical codes are difficult to implement due to the lack of any enforcement mechanism. Similarly, professional standards are not easy to implement because they are self-imposed codes of conduct. Using the FET as a case, this book illustrates the implementation of the standards in developing and validating an English language proficiency test. Though the FET is not a test of extremely high

stakes, the case analysis has demonstrated the usefulness of professional standards for ensuring and improving the quality of a language test.

It is a pity that the book is written only in Chinese, as required by the funding body of the project, making it difficult to reach a wider audience outside of China. Nonetheless, the book is well-written and easy to follow for Chinese readers. I would recommend it for practitioners involved in routine operations of language test development and validation, researchers who are keen to improve the quality of testing services, and key stakeholders such as language teachers, learners, educational policymakers, and test users. It is worth noting that standards need to be updated on a regular basis so as to better perform their functions. The outbreak of the COVID-19 pandemic, for example, has had a significant and continuing impact on language testing practices. Online teaching and learning has become an imperative. More importantly, the construct of a language test needs to be re-defined due to the changing nature of language communication. New standards should therefore be established to provide guidance on innovative task design that allows for "more interactivity, multimodality, and more authentic representations of today's communication and target language use domains" (Isbell & Kremmel, 2020: 616).

Reviewed by Yan Jin
Shanghai Jiao Tong University

References

AERA, APA, & NCME. (1985). *Standards for Educational and Psychological Testing*. AERA.

AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing.* AERA.

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing.* AERA.

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

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

Boyd, K. & Davies, A. (2002). Doctor's orders for language testers: The origin and purpose of ethical codes. *Language Testing, 19*(3), 296-322. https://doi.org/10.1191/0265532202lt231oa

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

Davies, A. (2008). Ethics, professionalism, rights and codes. In Shohamy, E. & Hornberger, N. H. (Eds.), *Encyclopedia of Language and Education* (2nd ed.), Vol. 7, pp. 429-433. Springer. https://doi.org/10.1007/978-0-387-30424-3_191

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

De Jong, J. & Zheng, Y. (2011). *Research note: applying EALTA Guidelines – a practical case study on Pearson Test of English Academic*. Pearson.

Isbell, D. R. & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37*(4), 600-619. https://doi.org/10.1177/0265532220943483

Jin, Y. & Fan, J. (2015). Code of Ethics in Language Testing: What can Language Testing in China Learn from the ILTA Code of Ethics? In *Sociology of Language Testing*, Yang, H. & Gui, S. (Eds). pp. 115-131. Shanghai Foreign Language Education Press.