McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, Justice, and Language Assessment: The Role of Measurement.* Oxford University Press.

Reviewer: Gladys Jean, Université du Québec à Montréal, Montréal, Québec, Canada

Language tests are used to make decisions that can have important repercussions on someone's life and career. While a lot has been said in second language testing about internal measures of reliability and fidelity, these measures do not fully ensure that tests are just and fair. The issue of justice, which is related to the social role and consequences of assessment and testing, is central to the argumentation in this book, as is fairness which is the success with which a test can make valid inferences and where test scores are defendable. The authors, McNamara, Knoch and Fan, argue that it is essential to work at making language tests fairer and more just, and they make a compelling case for the use of Rasch analysis models for reaching this goal.

The introduction of the book problematizes very well the issues of fairness and justice in language testing by providing an example of the consequences of test scores on judging the ability of a professional to perform his or her job adequately and safely. It further addresses the more general question of how language tests are defendable for the uses to which they are put.

Chapter 2 discusses the guiding issue and theory of validity in relation to the two concepts put forward by the authors, justice and fairness of tests, and briefly introduces the potential of Rash measurements for looking into these concepts from a moral standpoint rather than from a mostly technical one. It brings forward important issues such as the artificial nature of tasks used in tests, the difficulty to differentiate language proficiency and technical or clinical knowledge, the use of language tests restricted to non-native speakers (although some native speakers may be less proficient in some aspects), the naive faith in tests and their results, and the defensibility of requirements for immigration and citizenship and how they are measured in tests.

The proceeding chapters present Rasch measurement from simple to more sophisticated analyses, using real language test data. Chapter 3, an introductory chapter to the basic Rasch model, explains how Rasch measurement can overcome the limitations associated with the classical measurements of item difficulty and item discrimination, test reliability, and person ability, and how it can detect issues of fairness (or unfairness) associated with them. As an example, Rasch measurement analyzes data to determine if the same raw scores in a set of data for two or more candidates bring the same meaning regarding their individual abilities. Rasch measurements can also compare the range of a group's abilities with the range of item difficulty and thus provide insights into the test's level of difficulty. Another useful feature of Rasch measurement is its ability to identify misfitting items which is an indication of poorly written test items or items that do not belong to a particular measurement trait.

Chapter 4 presents two extensions of the Rash Model: the Andrich rating scale model which is designed to examine semantic differential scale with scores along a continuum (e.g., from strong to weak for the assessment of vocabulary) and the partial credit model, conceived for extended response as in the case of speaking or writing tasks using semantic differential rating scales. Again, the authors argue that using Rasch analysis can provide answers otherwise not obtainable through traditional analyses,

answers based on fairness which may lead to fairer decisions. However, rating scale data analyses do not take into consideration the influence of the rater.

In fact, while analyses described so far took into consideration only two facets, person and item, the many-facets Rasch model described in Chapter 5, as its name indicates, can handle many facets. The analyses obtained through this measurement are useful to model rater effects (in this case with the help of the FACETS software program), some of these being not only indicators of consensus or consistency in pairs of raters as for traditional analyses, but also other facets such as rater leniency or harshness, candidate abilities, topic difficulty, use of an interlocutor, raters' characteristics, attitudes and training, rating scale, criteria/band descriptors, task characteristics, and test-taker individual variables.

To further exemplify the potential of Rasch measurement for fairness in language assessment, Chapter 6 presents studies that investigated fairness in the field of language assessment using this model, which have increased in number and scope in recent years. Examples of current issues presented in these studies include rater effects in performance language assessment, raters' language background, technical quality of measuring instruments (e.g., validity of the *Vocabulary Size Test* and the *Peabody Picture Vocabulary Test*), and performance of different subgroups due to potential bias in test items, test content, and test tasks.

Although until this last chapter the book's content has been quite accessible to the non-specialist, the last two chapters are particularly directed to users of Rasch measurement for their assessment programs, as well as to readers interested in learning how to apply it to larger sets of data. Topics in Chapter 7 include setting up data (sample size and missing data, for example), identifying initial issues such as test equating, differential item functioning, item banking and computer-adaptative testing, and finally reporting. Chapter 8 furthers this discussion by explaining thoroughly the thinking behind Rasch measurement so that users can get a better understanding of its applications, know how it compares to other types of language testing measurements, and understand the debates over its uses.

The book concludes by coming back to the central issues of the book, namely, justice and fairness, reminding readers that the use of a test could be justified, hence just, but because of its design weaknesses may not be fair. The opposite is also true: a well-designed test, thus far, could not be justified or well adapted to the context, and thus unjust. Several examples of the latter situation are given where the well-designed test is not adapted to the situation, for example in the case of citizenship tests.

Overall, the book is quite accessible to non-specialists interested in understanding language testing, and to professionals producing language tests or analyzing data from these tests. It presents a rather large number of examples of statistics that are always clearly explained, without going into extraneous details. The numerous connections made between topics and reminders of issues already discussed help readers understand the quite technical aspects of language test data. For readers more versed in the topic, however, these links and reminders could be felt as unnecessary repetitions. Because of its pedagogical approach, the book could make a good reader in a university course on language testing, especially because it suggests activities on a companion website using data files discussed in the chapters.