T. McNamara, U. Knoch & J. Fan. *Fairness, Justice, and Language Assessment*. Oxford, UK: Oxford University Press. 2019. Pp. 224.

I was on a gondola going to the top of the Alps when I had one of my first conversations with Tim McNamara. We were in Innsbruck, Austria, at the 2012 European Association for Language Testing and Assessment (EALTA) conference; he was a plenary-speaker, and I was a young scholar who had used his book *Measuring Second Language Performance* (McNamara, 1996) to guide my doctoral work. I nervously approached him and blurted that I was introduced to his work in an educational measurement class. The professor of the class had handed us a photocopied book chapter and extolled that the best explanation he had ever found on many-facets Rasch measurement (MFRM) was written by an applied linguist. As the only linguist in that cohort, my ears perked and I decided to search out a copy to buy for my personal library and found a used copy listed with the jaw-dropping price tag of over a thousand dollars. Since the book was out of print and in demand, I had to resort to an interlibrary loan to see the full version and ended up using it extensively for my dissertation. As we watched the city get smaller and smaller, I thanked Tim for his excellent explanation of concepts that can be quite complex. Tim was gracious, kind, and supportive; he mentioned that he had been thinking about revising and re-releasing that book. So, this summer when I was asked to review *Fairness, Justice, and Language Assessment* (McNamara, Knoch & Fan, 2019), I was thrilled to read that it was an expansion, extension, and update of the original book.

The book begins and ends by discussing and differentiating the aspects of *fairness* and *justice*. The authors define *fairness* as the extent to which an assessment actually measures what it claims to measure. It is an internal quality of a test and can be determined through rigorous analytical methods; the book demonstrates "the most useful of these for communicative assessments of spoken and written language, i.e. Rasch measurement" (p. 3). They define *justice*, on the other hand, as the way test scores are externally used. So, while tests need to be *fair* in order to be *just*, it is also possible for a test that is *fair* to be misused externally and thus *unjust*; it might be *just* for one purpose (e.g. determining readiness for college admittance) but *unjust* for another (e.g. verifying language ability as a work requirement for immigrant populations). The bulk of the book details how Rasch measurement can help test developers create fair assessments.

Chapter 1 introduces *justice* and *fairness* in language assessment and provides examples of examinees for whom these issues had real world consequences. Chapter 2 more fully delves into the relationship between validity, justice, and *fairness* and how "large-scale language tests are primarily administrative instruments serving bureaucratic purposes" (p. 10); thus, it behooves test developers to consider what is at stake with the tests they

create. While it is difficult to control how a test is ultimately used, the least test developers can do is ensure that the instruments they create are fair, and document the development in such a way as to minimise unjust uses of the instruments. Rasch measurement is particularly well-suited to investigate the fairness of language tests since it can not only analyze tests with selected responses, such as multiple-choice reading and listening exams, but it can also analyze tests of productive responses such as writing and speaking and account for rating variance that comes from human judges.

Chapter 3 starts the bulk of the book's content by introducing "the basic concepts and procedures of Rasch measurement in a way that should be clear and accessible for the non-expert" (p. 22). The authors further explain that "no previous statistical or mathematical knowledge is expected" (p. 22) and that they will use a simple dichotomous data set (i.e. answers are either right or wrong) to illustrate the difference between a classical test theory analysis and an analysis using the basic Rasch model. One feature that the revised edition maintains is a display of raw data sorted in different orders so readers can easily see the connection between Rasch measures and the scores for both people and items. The authors then (1) conduct a Rasch analysis of a reading test in Winsteps (Linacre, 2019), (2) provide a simple description of the procedures, and (3) help the readers interpret the results. Included in the chapter are explanations of how to interpret item/person maps (Wright maps), person ability, item difficulty, fit statistics, item and person summary statistics, and reliability. One thing to note is that the data used in the book is available for download from Oxford University Press, so readers can run the analysis themselves and compare their findings with what is in the book. I've been using Winsteps for a number of years and by following the tutorial from the website, I learned some new ways to do my analysis.

Chapter 4 moves beyond analyzing dichotomous data to explore Rasch analyses with polytomous data including the *rating scale model* (RSM) and the *partial credit model* (PCM). The RSM is used when all the items in a particular instrument utilise the same rating scale (e.g., 5-point Likert scale on a questionnaire), while a PCM is used when each question can have different point values (e.g., open-ended test questions). This chapter discusses the two models in depth by detailing what information they can provide, when they are used, the data types that are typically analyzed, and the differences among all the models. Then, similar to Chapter 3, the authors use Winsteps to conduct a PCM analysis of a listening test that permits partial credit scoring and explains how to interpret item/person maps (Wright maps), person ability and fit, item difficulty and fit, and rating scale analyses. The chapter concludes with a demonstration of the RSM by conducting an analysis of a questionnaire on reading and then provides an explanation on the

interpretation of the same output tables listed above. The data used in this chapter is also available for readers to download and analyze themselves.

Chapter 5 introduces raters, ratings, and the many-facets Rasch model (MFRM). Pervasive in any judgment of human performance is the effect of rater variability. As the authors note, "arbitrary and irrelevant rater differences…[, which] pose a threat to the validity of the score interpretations we make, and thus constitute a risk to test fairness" (p. 91). The chapter starts by discussing rater variability and the steps that are typically practiced to minimise rater differences and then observes that despite those best practices, differences among raters will always persist. The authors address how those differences emerge in divergent ratings, the inherent limitations in using traditional approaches, and what is at stake for examinees whose ability is near cut score levels. The authors then introduce MFRM and how it can be used to model rater effects to calculate a fair average that takes into account rater tendencies of leniency or harshness. The authors then introduce readers to the Facets software (Linacre, 2019), provide data from a writing test that has been rated, and conduct a Facets analysis first using the RSM for a holistic analysis and then the PCM for an analytic analysis. The authors explain how to interpret the Wright map; the measurement reports on candidates, raters, and criteria (traits); and rating scale diagnostics. The chapter concludes by exploring one of the strengths of MFRM—the ability to model interactions to detect bias—which is an important aspect to investigate when trying to create fair assessments. As with previous chapters, the datasets are available for the reader to replicate the analyses presented.

Chapter 6 delves into the question of "how, specifically, do we investigate fairness issues in assessment using the tools of Rasch analysis?" (p. 129). The chapter then surveys scholarly work that has investigated fairness using Rasch measurement, including different academic journals in which studies have been published and the types of fairness issues examined including rater effects (variability, behavior over time, language background), the quality of the test instrument, test method and performance, and exploring the performance of different subgroups.

Chapter 7 moves beyond the basics of Rasch measurement and explores four different topics: "(1) setting up the data, (2) identifying initial problems, (3) further investigations and issues, and (4) reporting" (p. 146). In the first section, the authors address the effect that sample size and missing data have on the analyses conducted, and provide information on creating rating designs (particularly helpful when setting up MFRM studies in which connected subsets between facets are vital), and the use of anchor items. The second section helps the readers identify if their data violates any of the assumptions needed to carry out a Rasch analysis, namely unidimensionality and local independence.

The third section explores the issues of test equating (i.e. how to link several tests or forms to the same scale) and Differential Item Functioning (DIF) (i.e. how to detect items that potentially have bias to subgroups of examinees). The final section explores reporting test results and discusses: (1) "converting logits to meaningful units and rescaling" (p. 162), (2) "mapping abilities and change over time" (p. 163), and (3) "scaling and change in ability over time" (p. 165).

Chapter 8, titled "Data, Models and Dimensions," explores the "conceptual issues underlying the applications of Rasch modelling to language test data" (p. 167). The authors first provide more depth and historical context for the Rasch family of models (Georg Rasch's inception of the basic Rasch model, David Andrich's work with the rating scale model, Geoff Master's development of the partial credit model, and Mike Linacre's creation of the many-facets Rasch model). The chapter then addresses some of the debates within the educational measurement community on the use of the Rasch model, in particular by comparing it to other IRT models (e.g. 2PL, 3PL) that allow for the additional parameters of discrimination and guessing. While the inclusion of additional parameters may better model data for any given test, these models lack the property of objective measurement which puts both people and items on a common metric. Another criticism of all IRT models (including Rasch) is that language with its inherent complexity cannot be considered unidimensional and thus violates one of the underlying assumptions of the analysis. The authors contend that there is a difference between types of unidimensionality—psychological and psychometric—and that if evidence of psychometric unidimensionality can be established, then measurement is possible. Finally, the chapter compares Rasch to the data analytic models of G-theory and structural equation modelling and then evaluates strengths and weaknesses found therein.

Chapter 9 concludes by reconciling *fairness* and *justice*. The authors once again argue that expertise in sound measurement principles are a prerequisite for test developers to create high quality assessments that are fair and measure what they purport to measure. Once the quality of the test has been established for a specific purpose, then all other uses of the test should be evaluated. "Tests do not exist in a social and policy vacuum but always serve social goals and embody values that are often not made fully explicit. The values that the tests embody and policies that they serve are often not within the control of test developers and researchers; however, even if we cannot alter them, we do have a responsibility to understand them" (p. 191). The authors then share some of their experiences regarding *fairness* and *justice* in language testing.

The book is well-written and easy to follow. In fact, as I was preparing the syllabus for my graduate-level testing class, I decided to include the book as one of their elective options. I always have a few students who will use Rasch analyses for theses and projects, and this is an outstanding resource for them. My only criticism concerns the title itself. There is nothing contained therein to indicate that the book serves as a primer on Rasch measurement and provides exercises to help readers learn software to conduct those analyses. In fact, I had the book sitting in my office for close to a month before I opened it up and realised this was the long-awaited revision that Tim mentioned all those years ago. While we have the adage "never judge a book by its cover," it is human nature to do so, and I worry that since the title contains no reference to the psychometric procedures needed to measure language, many may not even look at it. But, I can highly recommend the book for anyone that wants to familiarise themselves with the principles of Rasch measurement but have heretofore been hesitant because the topic seems too technical.

Reviewed by Troy L. Cox

Brigham Young University

References

Linacre, J. M. (2019). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2019). Facets computer program for many-facet Rasch measurement, version 3.82.0. Beaverton, Oregon: Winsteps.com

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.