PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2014, 51(2), 168-171

Pontificia Universidad Católica de Chile
Facultad de Educación

Book Review

A review of the third edition of "Test Equating, Scaling, and Linking: Methods and Practices"

Fernanda Gándara

University of Massachusetts Amherst, USA

Kolen, Michael J., & Brennan, Robert L. (2014). **Test Equating, Scaling, and Linking: Methods and Practices**, 3rd edition, New York: Springer, 566 pages.

This year, the third edition of the book *Test Equating, Scaling, and Linking: Methods and Practices* by Michael J. Kolen and Robert L. Brennan (2014), was published. The book is intended to be used in psychometrics instruction and applied testing settings. Its purpose is to equip the reader with the principles of *equating*, *scaling*, and *linking*. Overall, the authors do an outstanding job in explaining and structuring the content, which flows logically and at a reasonable speed. However, in order to fully profit from the text, the reader should possess some knowledge of calculus, as well as advanced knowledge of Classical Test Theory (CTT) and Item Response Theory (IRT). The novice reader can refer to Crocker and Algina (1986) for the topic of CTT, and Hambleton et al. (1991) for IRT. They may also refer to Muñiz (1994, 1990) for a Spanish treatment of these topics.

Post to:
Fernanda Gándara
University of Massachusetts Amherst, USA
Hills South, 111 Thatcher Rd, Office 168, Amherst, MA, 01003, USA
Email: mgandara@educ.umass.edu

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi: 10.7764/PEL.51.2.2014.26

The structure and content of the book are very similar to that of the second edition, which also included extensive coverage of scaling and linking methods. Chapter 1 presents an overview of the terminology and methods upon which the rest of the book is based. Essential to the reader who is unfamiliar with the technicalities related to equating, this chapter refers to (a) the conceptual differences between equating, scaling, and linking, (b) the five properties that equating should possess, and (c) the different data collection designs that can be used in an equating study. Chapters 2-6 present different equating methods, under certain data collection designs. Chapter 2 refers to the observed score equating methods under the random groups design. The authors cover in detail the procedures to conduct linear and equipercentile equating under such a data collection design. Brief but useful considerations are presented in regard to the implementation of these methods under the single-group design. Chapter 3 explains the rationales and methods related to smoothing equated scores using equipercentile equating. Smoothing methods correct irregularities in the sample distributions to improve the equating results. The authors describe two presmoothing methods: (a) log-linear methods for observed scores and (b) the beta4 method for true scores. They also describe the use of splines in postsmoothing, and refer briefly to the kernel method of equating, which uses presmoothing methods. Particularly important is the discussion of the relationship between smoothing methods and equating error. Chapter 4 refers to the linear methods under the non-equivalent groups design. The authors describe the Tucker and Levine methods for observed scores and the Levine true score method. In this edition, they extend the comparisons between the methods and the discussions of applications and related topics. Chapter 5 refers to equipercentile methods for the non-equivalent groups data collection design. The content is structured differently from the second edition, by grouping methods into two categories: (a) frequency estimation methods, and (b) other methods—including the chained equipercentile equating. Chapter 6 covers the topic of IRT methods of equating and applications that are unique to this context. The authors review the IRT concepts, models—both dichotomous and polytomous—and assumptions, that are needed to understand the methods. They also explain the logic of IRT equating under different data collection designs. They describe IRT observed scores equating methods, including: (a) mean/mean, (b) mean/sigma, (c) Haebara, and (d) Stocking and Lord. The authors then describe IRT true score equating and its application to observed scores. As in previous chapters, the authors provide comparisons of the methods and rationales for choosing between them, when appropriate. A brief reference to Rasch equating is provided. Lastly, the authors describe these methods in the context of polytomous models, with a similar level of detail.

Chapter 7 remains essentially unchanged from that of the previous edition, and provides a detailed treatment of the topic of standard errors of equating. In particular, on estimating random error (as opposed to systematic error). Standard error of equating is the standard deviation of the equated scores across hypothetical replications at each score point. The authors introduce two methods for its estimation: (a) bootstrapping methods—including parametric bootstrapping, and (b) several variations of the delta method—which is an analytic procedure based on Taylor's expansion series. The most interesting part of the chapter is that pertaining to some practical applications of these concepts, such as the determination of sample sizes for ensuring that the standard error of equating is less than a given quantity. Chapter 8 discusses the practical issues pertaining to equating, which are essential to the practitioner. This is the chapter where the fundamental concepts and many unanswered questions concerning the equating methods and designs, fall into place. The authors refer in depth to the essential considerations in choosing among data collection designs, relative to administration and test development conditions, statistical assumptions, and effects. These include considerations about population and sample sizes. A similar analysis is provided for the choice of statistical procedures or equating methods, and for choosing among results when different methods are applied. A step-by-step guide to perform quality control checks on the data is also provided. The last part of the chapter refers to issues that may threaten score comparability, which is relevant given that assessments are increasingly delivered in multiple administration modes and versions. In this edition, a broader section about constructed-response and mixed-format tests includes the content previously contained in the section about comparability of performance assessments.

Chapter 9 gives extensive coverage to the topic of scaling. The first two sections are new to this edition. The first section refers to basic terminology, and brings attention to the complexity of scaling. The second section discusses issues pertaining to scaling scores on mixed format tests, such as the different weighting decisions for items of different types and their effects on the psychometric properties of the test. The rest of the chapter is structured similarly to the previous edition. The authors refer to topics such as the transformations that may be applied on scaled scores, their rationales and implications, or the rationales and methods to introduce normative information. A brief subsection on how to determine the number

of scale points is particularly interesting to those creating new testing programs. The authors also refer to ways to incorporate content information, and to the topic of composite scores arising from testing batteries. The last part of the chapter deals with the topic of vertical scaling: (a) designs, (b) methods and the comparison of their results, (c) maintenance, (d) research around vertical scaling, and (e) considerations about growth models. Chapter 10 covers the topic of linking, addressing the terminology and conceptual frameworks. In this edition, in addition to the Mislevy/Linn framework, the authors include the Holland and Dorans framework. The last part of the chapter is dedicated to the topic of invariance, largely based on the work of Dorans and Holland (2000). The authors extend the concept, including more statistics to evaluate group invariance and their multiple considerations. The authors stress the importance of building scales that facilitate interpretation of scores and minimizes their misinterpretation and therefore, misuses.

This edition did not include Chapter 11, which was the least informative of the second edition of the book (Eignor, 2006). For further information about the previous edition, which is very similar, the reader can refer to Eignor (2006), Skaggs (2006), or Chiu, Carr, & Li (2007).

The book is clearly useful for instruction. The authors spend a large part of the text demonstrating the connections between theory and assumptions for the different methods and designs. They support their explanations with detailed examples that the reader can easily replicate. The explicit references to the corresponding software allow direct application of the knowledge to the datasets of interest. And they include a large list of answered exercises in each chapter.

The book is also well suited for applied settings. It covers all the important information that has been published in the literature to date. It provides thorough discussion of the topics that concern practitioners the most, particularly in Chapter 8. And the focus of this edition has slightly changed towards the inclusion of topics that are becoming increasingly important in the field, such as issues related to mixed-format tests or growth models.

The book could be improved in some ways. One suggestion would be to extend the topic of scaling. In particular, the topics of vertical scaling and of growth models deserve a chapter in itself. Another suggestion would be to include more policy considerations in some of the discussions, when appropriate. Many of the current large-scale testing programs include national, state, or international assessments used for policy decisions. Referring to the implications that the topics of equating, scaling, or linking may have in these contexts may increase awareness of their importance and reduce certain misunderstandings that often exist in score interpretation and use of scores.

This book is certain an essential reference for the reader interested in psychometrics. While in-depth discussions about certain advanced topics within equating can be found in other texts (see von Davier, Holland, & Thayer, 2004; Dorans, Pommerich, & Holland, 2007; or von Davier, 2011), this book remains as the fundamental guide to equating, scaling, and linking.

References

Chiu, C., Carr, P., & Li, I. (2007). A review of "Test Equating, Scaling, and Linking: Methods and practices". *Journal of Educational and Behavioral Statistics,32*(2), 223-228.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, Fl.: Harcourt Brace Jovanovich.

Dorans, N.J., & Holland, P.W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales.* New York, NY: Springer.

Eignor, D. R. (2006). Test equating, scaling, and linking methods and practices, 2nd edition [Book review]. *Journal of Educational Measurement,43*(2), 169-172.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Fundamentas of item response theory.* Newbury Park, CA: Sage.

Muñiz, J. (1990). *Teoría de respuesta a los ítems.* Madrid: Pirámide.

Muñiz, J. (1994). *Teoría clásica de los tests.* Madrid: Pirámide.

Skaggs, G. (2006). Book review: Test equating, scaling, and linking (2nd ed.). *Applied Psychological Measurement, 30*(6), 511-513.

Von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York, NY: Springer-Verlag.