Book Review

Applied Psychological Measurement
2015, Vol. 39(7) 575–578
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
apm.sagepub.com

Cheng, Y., & Chang, H.-H. (2014). *Advancing methodologies to support both summative and formative assessments*. Charlotte, NC: Information Age Publishing. 333 pp. \$45.99. ISBN 978-1-62396-595-2

Reviewed by: Justin L. Kern, *University of Illinois at Urbana-Champaign, USA*.

DOI: 10.1177/0146621615595498

As national standards in education have gained steam as an engine for economic growth in the United States, the importance of gauging the progress of students has risen drastically at the national level, the state level, and the local level for accountability purposes. As such, tests now inherently play a larger role in the educational process than ever before. Although there is a significant debate about how that information is to be used (and how to enforce accountability), everyone can agree that no matter the usage, the information should be as accurate as possible. At the same time, there is the recognition that administering tests is costly and time-consuming, so test administration should have the dual goals of maximizing accuracy and minimizing costs; modern test construction and administration techniques have been developed to handle both.

The role that teachers play in our educational system has become increasingly difficult to navigate as national standards have been implemented. With the standards, teachers are being called on to be held accountable for the progress—or lack of progress—of their students. With an increased scrutiny and higher accountability, teachers are in need of tools to help them with their goals regarding students. Currently, there is a movement within the educational testing field to create methodologies designed to help inform teachers of the skills that students possess (or do not possess) to assist teachers in determining the areas that students need intervention. In particular, cognitive diagnostic models are being developed for this purpose.

Clearly, the current educational climate underpins the need for a volume informing test users and researchers of current advances in methods of obtaining more information more efficiently both for summative and for formative assessments. The edited book *Advancing Methodologies to Support Both Summative and Formative Assessments* neatly fits in this gap.

This book is organized into four sections. The first section contains three chapters on topics for making tests more efficient using modern tailored testing designs. The second section contains three chapters on obtaining more information from and about test items. The final two sections contain four chapters each about gaining diagnostic information via subtests and subscores (using multidimensional models), and via cognitive diagnostic models, respectively.

Chapter 1 is focused on methods of automated test assembly. These include combinatorial optimization techniques, heuristic assembly techniques, and Monte Carlo techniques. The chapter does well in explaining the strengths and weaknesses of each method so that the practitioner may make the best decision for his own test. The author further goes on to discuss the practical issues that are related to test assembly, such as feasibility of the test constraints, the number of test forms, and developing item pools.

Chapter 2 introduces multistage testing (MST), including historical and modern developments. The author discusses the parts of MST, such as test design, test assembly, and routing rules. Many practical suggestions are given in the chapter, along with a nice summary of the research on MST to date. This information will prepare practitioners and researchers alike for their respective goals. Finally, a new method for test assembly, on-the-fly MST (OMST; Zheng

and Chang, 2011, 2015), in MST is introduced. This method bypasses the need for up-front form/panel assembly using a computer adaptive testing (CAT) algorithm to assemble a form after the routing test is completed; it is clearly more flexible as well (an infinite number of routes are able to be obtained). The pros and cons of this method are given. Finally, a short discussion of important details to help a practitioner choose between using MST, CAT, and OMST is given.

In Chapter 3, we are treated to a discussion of repeated statistical tests in group sequentially clinical trial designs. The first part of this chapter introduces terminology common in clinical trials, followed by an introduction to several types of sequential testing procedures. This introduction is important for researchers interested in sequential testing designs (such as MST), but it may be a bit dense for practitioners trying to make decisions about their own tests.

Chapter 4 introduces the concept of online calibration, the goal of which is to calibrate new items with the help of already calibrated operational items in a test. It gives an overview of all of the necessary elements of online calibration and continues with three examples of online calibration designs. It concludes with problems currently faced in this research area.

Chapter 5 is a discussion about issues surrounding the error in estimating (calibrating) item parameters. Despite item parameters only being estimates, they are often treated as though they are the true parameters; this can potentially be harmful. This chapter discusses several domains within measurement that can be affected by this error, as well as how the effects of estimation error can be assessed.

When using mixed-type tests (i.e., tests with both dichotomous and polytomous items), ability estimation is more efficient if the polytomous items are more highly weighted. Using fixed weights in CAT, however, may not be an optimal estimation strategy. In Chapter 6, the author discusses several strategies for allowing for an optimal weighting scheme at stage of an adaptive test, and includes a real data example. The chapter concludes with a discussion of the pros and cons of using these techniques.

Chapter 7 introduces multidimensional item response theory (MIRT) along with many of the associated models. The main topic of this chapter is on the assessment of the dimensionality of a test. There is a lengthy review of the pros and cons of many different dimensionality assessment procedures, as well as a discussion of assessing the dimensionality of MST and adaptive tests—situations when there can be high levels of missing data by design.

With the shift toward an increased use of formative assessments in elementary and secondary education, there has been a heightened demand for making scores and subscores more reliable and accurate. Chapter 8 discusses using MIRT for the reporting of scores. Much of the chapter is devoted to different ability estimation routines and the comparisons between them in terms of reliability and accuracy. There is also a lengthy discussion about linking scores in MIRT.

Chapter 9 continues the theme of examining MIRT models, but now in the context of CAT. Multidimensional adaptive tests have many advantages over standard unidimensional adaptive tests, but with these advantages come additional challenges. These challenges include ability estimation, item selection, and stopping criteria. In particular, this chapter devotes a great deal of space to the topics of item selection and stopping criteria, and the issues surrounding them. The chapter concludes with a discussion of future directions in multidimensional adaptive testing.

Chapter 10 discusses Rasch models and their multidimensional extensions. Rasch models have a long tradition of usage in the educational and social sciences, largely because of their simplicity in application and interpretation. Much of this chapter is dedicated to presenting a series of applications of multidimensional Rasch models for practical modeling situations, such as bifactor modeling and rater-effect modeling.

Chapter 11 introduces cognitive diagnostic CAT (CD-CAT; Cheng, 2009) in the context of a real large-scale implementation in China. The goal of this chapter is to show that CD-CAT has great promise for large-scale implementation. It goes through all of the necessary steps from developing an item bank and constructing a Q-matrix to implementing item selection to field testing and validation. The chapter concludes with a discussion of further research directions.

Chapter 12 extends the topic of cognitive diagnostic assessments to an application of the fusion model (Hartz, Roussos, & Stout, 2002). It begins with a historical development of the fusion model from the unified model, followed by a step-by-step description of an application of the model to state assessment data. The chapter concludes with a simulation study based on the results of the application.

Chapter 13 is the second chapter in the book to focus on online calibration. In this chapter, however, the online calibration is paired with the cognitive diagnostic model. The bulk of the chapter is devoted to the test design to be used and the methods for use in item parameter estimation; these are generally straightforward extensions of those used in the online calibration of an IRT model. A great addition to the topic specifically for the CD model is the online calibration of the Q-matrix for the new items. For this, the author proposes to use what they term *the joint estimation algorithm* (Chen, Xin, Ding, & Chang, 2011).

The final chapter of this book deals with validation of cognitive diagnostic measures via person-fit statistics. After an introduction to the current state of person-fit measures, the chapter has a review of person-fit studies for cognitive diagnostic tests organized by proposed fit indices. These indices include the hierarchy consistency index, the likelihood ratio test of person fit, and the response conformity index. The chapter ends with a call for increased focus on person-fit analysis in cognitive diagnostic assessment research.

This goal of this book, in the words of the editors, is “to provide state-of-the-art coverage on new methodologies to support traditional summative assessment, and more importantly, for emerging formative assessments.” In this, the book succeeds. Not only does it cover a wide array of topics integral to modern assessment, but it does so with clarity. None of the chapters is overly verbose, although the amount of information packed into them is immense.

The biggest strength of this book is that it has brought together many diverse methodologies, all with the singular goal of improving assessment, into one volume. Of great value are the real data examples. By sharing these, the individual authors show that these new sophisticated methods are not just shiny new toys, but can have real impact on assessment. Of greatest importance is the final section (Chapters 11-14) of the book focusing on formative assessment. As far as the reviewer can tell, there are no other examples of books in the literature with large sections devoted to models that focus on formative assessment. In particular, Chapter 11 shows that with the right level of care, that CD-CAT can be implemented to great use. Ultimately, as articulated by Chang (2015), individualized instruction of students by their teachers would be aided greatly with an implementation of CAT technology for this purpose.

In the reviewer's opinion, this book would be useful for researchers and practitioners alike. For the researcher, the information in this book is quite valuable as a tool to get the mental juices flowing and inspire new ideas for research. Many topics covered in the book are on the cutting edge. For instance, the first article to introduce the topic of OMST was only recently published in *Applied Psychological Measurement* in 2015. Although this implies that much research should be done before OMSTs are implemented in practice, its inclusion in this volume shows that it has potential. For the practitioner, the book shows the utility of these new methods for use in large-scale implementation, particularly in the final four chapters. Throughout the book, every chapter argues that the methods proposed and discussed can (and, in many cases, should) be considered for use in real testing situations, which, ultimately, is the goal of every researcher. Now the practitioners just need to pick up this volume.

References

- Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.
- Chen, P., Xin, T., Ding, S., & Chang, H.-H. (2011, April). *Item replenishing in cognitive diagnostic computerized adaptive testing*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74, 619-632.
- Hartz, S., Roussos, L., & Stout, W. (2002). *Skill diagnosis: Theory and practice* [Computer software user manual for Arpeggio software]. Princeton, NJ: ETS.
- Zheng, Y., & Chang, H.-H. (2011, April). *Automatic on-the-fly assembly for computer adaptive multistage testing*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.
- Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39, 104-118.