Marquette University

e-Publications@Marquette

Psychology Faculty Research and Publications

Psychology, Department of

2018

Book Review of *The Basics of Item Response Theory Using R*

Brooke E. Magnus

Follow this and additional works at: https://epublications.marquette.edu/psych_fac

Part of the Psychology Commons

Marquette University

e-Publications@Marquette

Psychology Faculty Research and Publications/College of Arts and Sciences

***This paper is NOT THE PUBLISHED VERSION;* but the author's final, peer-reviewed manuscript.** The published version may be accessed by following the link in the citation below.

Measurement : Interdisciplinary Research and Perspectives, Vol. 16, No. 3 (2018): 203-207. DOI. This article is © Routledge Taylor & Francis Group and permission has been granted for this version to appear in e-Publications@Marquette. Routledge Taylor & Francis Group does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from Routledge Taylor & Francis Group.

Book Review of *The Basics of Item Response Theory Using R, edited by Baker, F.B., & Kim, S.H. (2017). New York: Springer. ISBN: 978-3-319-54204-1*

Brooke Magnus
Department of Psychology, Marquette University, Milwaukee, WI

Abstract

This article reviews the book *The Basics of Item Response Theory Using R* by Baker and Kim (2017). It describes the structure and goals of the book, provides an overview of each chapter, and concludes with general comments. Both strengths and limitations of the book are discussed.

Keywords

Item response theory; R

The Basics of Item Response Theory Using R is a practical and concise introductory text for anyone looking to learn the nontechnical foundations of item response theory (IRT) in the context of educational measurement. It is a follow-up edition to Frank Baker's two previous books, *The Basics of Item Response Theory*, published in 1985 and 2001, respectively. All three editions of the book are written in a tutorial style with plenty of numerical examples, designed to provide those with little experience in psychometrics and statistics an accessible introduction to the basic concepts of IRT. Like the previous two versions of the book, eight major topics are addressed over eight chapters: the item characteristic curve (ICC), ICC models (for dichotomous responses), parameter estimation, test characteristic curves (TCCs), ability estimation, the information function, test calibration, and specifying the characteristics of a test. As the authors acknowledge in the preface, the treatment of these topics remains largely unchanged from the 2001 version of the book. The major updates come in software instruction and implementation. Whereas the previous editions demonstrated each technique using Baker's own computer programs developed specifically for IRT—an Apple II computer program in 1985 and a Visual Basic 5.0 program in 2001—the tutorials in the current edition shift the focus to R, the freely available statistical software and programming language that is well on its way to becoming ubiquitous in quantitative social science research. The strong graphical capabilities of R make the software especially advantageous for IRT analysis, in which graphics very often form a critical component of results presentation. In light of the growth in popularity of R in the recent years, I view this book as a timely and welcome addition to the introductory IRT literature.

Goals and structure of the book

Rather than using any of the existing R packages developed specifically for IRT—e.g., mirt (Chalmers, [1]), ltm (Rizopoulos, [2])—the tutorials in this book require only those functions that are available in base R. I believe that this is one of the most invaluable features of the book: In learning how to program the relevant functions in R, beginners have the opportunity to develop a deeper understanding of the underlying theory of IRT, not just its software implementation. While no prior knowledge of R is assumed, a minimal understanding of object-oriented programming may be helpful. For those who have a more limited knowledge of R, the book also includes an appendix (Appendix A) that covers introductory R topics, including instructions for downloading the software, inputting data files, and accessing the help guide for any of the specific functions used in the book. It also illustrates the use of R for some elementary statistical methods that may be helpful for new users (e.g., t tests and simple linear regression). All of the R functions used in the book are included in a convenient R package (birtr), available for download through the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/birtr/index.html).

Each chapter begins with an introduction to one of eight general topics, providing some relevant history and just enough detail for the reader to understand the important concepts without being burdened with technical detail. The mathematics behind the theory is intentionally kept to a minimum. After presenting the basics of a concept, the authors then describe and demonstrate with code how the procedure can be implemented in R. In addition to providing the R code, they also describe what the reader should see on the screen as they follow along with the analyses, a feature that may be especially helpful to new R users. Several numerical and computer examples are interspersed throughout the text, and exercises at the end of each chapter give readers the opportunity to test their knowledge of both the concept and its implementation in R. Finally, like the previous editions of the book, each chapter concludes with a section entitled “Things to Notice,” enumerating several of the key points from the chapter that the authors believe are the most important takeaways for applied researchers and practitioners. The complexity of the topics, as well as the accompanying R code, builds with each chapter; thus, the readers are encouraged to return to the computational examples from previous chapters as needed.

Chapter overviews

Chapter 1 begins with a brief overview of ability as a latent variable and highlights the fundamental differences between classical and modern test theory. The authors quickly delve into a discussion of the ICC as the building block of IRT, primarily using graphics to convey the meaning of item difficulty and discrimination parameters within an educational measurement context. Deviating somewhat from discussions of the ICC or trace line typically seen in other introductory IRT textbooks, the authors describe the item parameters without ever referring to the logistic function, which they cover in a later chapter. Instead, they describe these parameters graphically and more generally as variables that can be manipulated to produce ICCs with varying locations and slopes, providing plain verbal descriptors to interpret the magnitudes of these parameters (e.g., a very easy item with low discrimination). This is where the authors first introduce the reader to writing one’s own general R functions. They introduce a few lines of R code for plotting a single ICC for a given set of item parameters, and then show how these lines of code can be wrapped into a function. The reader can then observe how the appearance of the ICC changes when different item parameters are swapped in and out of the function. Chapter 1 also discusses some of the important arguments of the “plot” function in base R, which are used throughout the remainder of the book.

Chapter 2 covers some of the most commonly used IRT models for dichotomous responses, including the two-parameter logistic (2PL), the three-parameter logistic (3PL), and the Rasch models. Building on the ICC tutorial from the previous chapter, the authors discuss how a logistic function, which they refer to more generally as an ICC function, underlies the ICC plot. They then describe how this function can be used to calculate the probability of a correct response across varying levels of ability. One of the exercises that readers may find particularly helpful involves using the logistic function to compute the probability of answering an item correctly at several different ability levels. The reader quickly learns that for the same set of item parameters, the probability of answering the item correctly increases with ability. The authors demonstrate similar computational exercises for the Rasch model, the 2PL model, and the 3PL model, showing how the process of computing these probabilities can be automated by writing an R function. This is a particular advantage of learning IRT by using R: By changing any of the item parameter arguments of the function, the reader can easily see the relationship between the item parameters and the ICCs. Applied researchers and practitioners are likely to find the table that interprets “typical” values of item parameters especially useful. While this chapter includes sufficient coverage of three commonly used IRT models for dichotomous responses, it does not address IRT models for ordinal or multinomial responses.

Whereas the first two chapters treat item parameters as known quantities, Chapter 3 raises the issue that parameter values are actually unknown and must be estimated from the item response data. For the sake of simplicity, the authors discuss item parameter estimation under the assumption that examinee ability scores are known. While the concepts in this chapter are a bit more mathematical than those of earlier chapters, the authors limit their coverage of maximum likelihood estimation to a more conceptual level and refrain from discussing any of its technical details (though for interested readers, they do include a footnote about the Newton-Raphson algorithm). The description of maximum likelihood estimation should be accessible to anyone who has had an introductory statistics course and is familiar with the concept of minimizing error between observed and model-predicted values. The authors expand on the concept of minimal squared error in their brief discussion of one method of evaluating item fit. One of the strengths of this chapter includes a section on the group invariance of item parameters, a feature that sets IRT apart from classical test theory (CTT). Using numerical examples with ICC plots, the authors show that within sampling error, item parameters are not dependent on the ability level of the particular sample. The clarity of their explanation is noteworthy, and I now refer students and colleagues to this chapter when they wish to better understand this particular strength of IRT. They accompany their explanation with R code that illustrates group invariance of item parameters under the Rasch model.

The first three chapters focus on the analysis of a single item; the concept of a test of multiple items is first introduced in Chapter 4. After describing the idea of "true score" within a CTT framework, the authors then extend the concept to the IRT framework, in which an examinee's true score is the expected proportion of items answered correctly. They then show how one can plot the expected proportion correct as a function of ability level, resulting in the TCC. Throughout their explanations, they emphasize that the overall form of the TCC depends on the particular combination of item characteristics used to build the test. The R functions become a bit more advanced in this chapter, requiring vectors of item parameters as opposed to scalars; however, because the logic and form of the functions are parallel to those discussed in earlier chapters (i.e., the function for the TCC requires vector arguments and the function for the ICC requires scalar arguments, but otherwise the lines of code are quite similar), the reader should not have much difficulty extending the R functions from the single-item to the multiple-item case.

The first four chapters focus on the characteristics of the item, assuming that examinee ability is known. In contrast, the focus of Chapter 5 is on characteristics of the examinee and ability estimation—specifically, ability estimation for a single examinee when item parameters are known. The authors choose to limit their discussion of ability estimation to an iterative, maximum likelihood-based procedure, likely due to the intentional brevity of the book. They also introduce the concept of a standard error to quantify the precision of the examinee's ability estimate. Notably, the authors refer to the two cases in which maximum likelihood ability estimates are undefined: when an examinee answers either no item correctly or all items correctly. Rather than mentioning any alternative estimation techniques that do not suffer from this limitation, however, they simply state that it is impossible to estimate an examinee's ability in these two scenarios, and that these individuals must be removed from the scoring sample or assigned some arbitrarily small (or large) value of ability. Indeed, this is what is done in their accompanying R code, in which extreme scores are assigned values of *±2log(number of items).* Due to the limited scope of the book, this chapter's sole focus on maximum likelihood scoring is not necessarily a drawback; however, I believe that acknowledgement of alternative scoring methods that do not have this problem would have been beneficial, even for the novice reader. Despite the potential limitation of its coverage of scoring, this chapter also has a number of strengths, including a very clear explanation and computer demonstration of the principle of item invariance in ability estimation. The R code for maximum likelihood estimation is also very well annotated, and it is a good resource for those wishing to understand how to implement similarly iterative procedures in R.

Chapter 6 introduces the information function, for both a single item and a test. The authors begin with the general vernacular definition of information and then transition into its more technical definition relating to the precision of a parameter estimate, emphasizing that within the IRT framework, information varies as a function of ability level. Because the information formulas are later used in the R code for constructing information curves, the authors present the item information formulas for the 2PL and 3PL models; however, as appropriate for the intended audience of the book, they do not show their derivations. For a given set of item parameters, a Microsoft Excel-like table of values is used to illustrate the computations of item and test information at different levels of ability for the 2PL and 3PL models. While this chapter provides ample R code for calculating and plotting item and test information curves, the authors also focus a great deal on interpretation, paying particular attention to the factors that influence the size and shape of the information function. Readers may find the "Things to Notice" section of this chapter particularly useful for its clear descriptions of how test information curves can vary under different conditions.

Interestingly, unlike the previous six chapters, which include examples of the 2PL and 3PL models in addition to the Rasch model, Chapter 7 focuses solely on the Rasch model. The authors illustrate test calibration under the Rasch model using a two-stage joint maximum likelihood estimation method, drawing attention to the issue of defining the ability scale metric. Because of its focus on the Rasch model, this chapter describes anchoring the ability scale metric by fixing all item discriminations to 1 and the average of the difficulty parameters to 0; the other common method of setting the ability scale by fixing its variance to 1 is not discussed. Three data examples of 10-item tests taken by the same set of 16 examinees are included, in which the tests are designed to be of easy, medium, and hard difficulty, respectively. For each data example, the authors provide R code for calibration and an interpretation of the resulting item difficulty and ability estimates. In Appendix C, which was part of Chapter 7 in the previous version of the book, the authors show how ability estimates from these three tests can be put on a common scale using equating methods. While this chapter includes a clear description of the estimation procedure and detailed R code, it only presents the case in which the summed score is a sufficient statistic for ability. For this reason, this chapter may have limited practicality for researchers and practitioners who routinely encounter data that require more complex measurement models. While this may be considered somewhat of a limitation, it is also unlikely that the technical details of item and test calibration are of great interest to the nonspecialist audience for whom this book is intended. For the interested reader, other well-documented R packages are available for this purpose (e.g., mirt, ltm).

The final chapter is a very brief introduction to specifying the characteristics of a test—that is, selecting items from a pool of precalibrated items. Three types of tests are described: screening tests, broad-range tests, and peaked tests. The chapter describes how one can choose items based on content and technical characteristics that meet the goals of the particular test. Detailing the procedure step by step, the authors then illustrate an example of choosing items for a screening test in which the goal is to maximize test information at a specific ability level that is considered a decision cut point. Aside from one copyediting mistake that may confuse some readers—a plot of a test characteristic curve that is captioned as a test information function—this chapter offers a clear and succinct example of how one can use information plots in R to help guide item selection. It ties in several of the most important topics from the earlier chapters (e.g., item parameter interpretation, ICCs, TCCs), serving as a nice overall conclusion to the book.

General comments

One of the primary strengths of this book is what I consider its "learn statistics by doing (and programming) statistics" approach. While there are several high-quality textbook options for learning IRT, few complement the explanations of the theory with such thorough instructions for software implementation. Moreover, I am unaware of any IRT textbook that uses R (or any other open-source program) as the accompanying software. I believe that its focus on R is one of the major advantages of this update, along with the many computational examples that are available for additional practice at the end of each chapter. Additional strengths of this book include its very accessible explanations of concepts that can often be difficult for new learners of IRT, especially those coming from a CTT background. In particular, I found the sections describing ability and item invariance to be quite clear. The use of technical jargon and mathematical formulas is kept to a minimum, and one does not need to be well-versed in psychometrics or statistics to follow the basic principles of IRT that are outlined in this book. Applied researchers and practitioners will likely also appreciate the focus on plain English interpretations of IRT models and parameters, as well as the heavy use of graphics throughout each chapter. Readers should be able to easily replicate any of the graphics for use in their own publications or technical reports.

As the authors acknowledge, this book is designed to provide readers with a working knowledge of IRT rather than the details of more specialized topics or cutting-edge techniques. I believe that *The Basics of Item Response Theory Using R* could serve as an excellent supplement for an introductory graduate-level course on educational measurement, or perhaps even an advanced undergraduate-level statistics course designed to expose students to the basics of programming in R. Because concepts are explained with no prerequisite knowledge of psychometrics or mathematics, the coverage of topics is necessarily limited in technicality. Similarly, the book's compromise between thoroughness and brevity means that many important topics of IRT are only tangentially covered (e.g., unidimensionality, model fit), or not covered at all (e.g., local independence, differential item functioning). Further, while the current edition of this book is novel in its presentation of R code, its coverage of concepts has remained largely unchanged since 2001. As a result, there are several places where it is obvious that the prose has not been updated to reflect the more current and widespread use of IRT. On a similar note, the book does not reference any of the more modern estimation methods (e.g., Bayesian techniques) or recent advancements in the field (e.g., multidimensional IRT). Last, because the book is written from an educational measurement perspective, it focuses only on models for responses that are scored dichotomously. Social scientists who regularly deal with Likert-type items and polytomous responses will likely need to consult with additional resources.

While such omissions may preclude this book from serving as a standalone text for a graduate-level IRT course, I believe that this edition offers great value for its R tutorials, its effective use of graphics, and its “need to know” approach to describing fairly complicated concepts. It will likely offer the greatest benefit to those who require only the essentials of IRT for their research and operations: educators and school employees who have perhaps never been exposed to IRT, practitioners working in the more applied divisions of testing programs, employees of government agencies, and other similar audiences. Because of its focus on R, it may also be useful to those who are primarily looking to learn the basics of programming in R, even those having a limited interest in its application to educational measurement. Rather than expecting a comprehensive treatment of IRT that includes the most current details of model assumptions, estimation, and scoring, readers should expect to learn only the fundamental and most important concepts, as the book’s title implies. While it will not turn a beginner into an IRT expert, *The Basics of Item Response Theory Using R* is an excellent starting place for readers who have some familiarity with CTT and are looking to expand their toolbox to include more modern techniques—in terms of both the method and the software.

References

1 Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48 (6), 1 - 29. doi: 10.18637/jss.v048.i06

2 Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. Journal of Statistical Software, 17 (5), 1 - 25. doi: 10.18637/jss.v017.i05