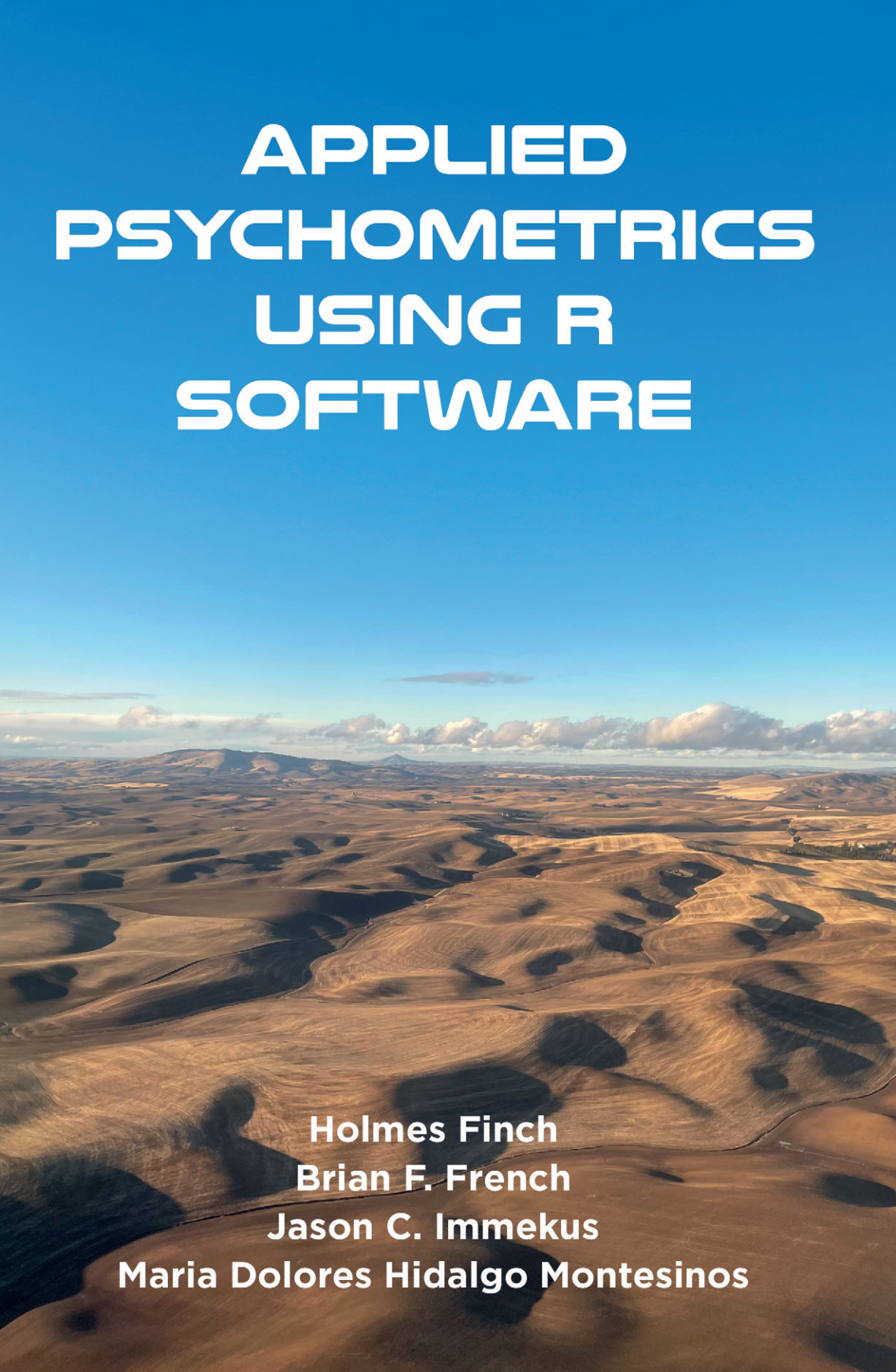
APPLIED PSYCHOMETRICS USING R SOFTWARE



Holmes Finch
Brian F. French
Jason C. Immekus
Maria Dolores Hidalgo Montesinos

Applied Psychometrics Using R Software

This page intentionally left blank

Applied Psychometrics Using R Software

By

Holmes Finch
Ball State University, USA

Brian F. French
Washington State University, USA

Jason C. Immekus
University of Louisville, USA

And

Maria Dolores Hidalgo Montesinos
University of Murcia, Spain



United Kingdom – North America – Japan
India – Malaysia – China

Emerald Publishing Limited, New Era House, 8 Oldlands Way, Bognor Regis,
West Sussex, England PO22 9NQ

First edition 2027

© 2027 Emerald Publishing Limited.
All rights of reproduction in any form reserved

Cover photo: Brian F. French

Reprints and permissions service
Contact: www.copyright.com

No part of this book may be reproduced, stored in a retrieval system, transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without either the prior written permission of the publisher or a licence permitting restricted copying issued in the UK by The Copyright Licensing Agency and in the USA by The Copyright Clearance Center. Any opinions expressed in the chapters are those of the authors. Whilst Emerald makes every effort to ensure the quality and accuracy of its content, Emerald makes no representation implied or otherwise, as to the chapters' suitability and application and disclaims any warranties, express or implied, to their use.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-80791-471-4 (Print hardback)
ISBN: 978-1-80791-473-8 (Print paperback)
ISBN: 978-1-80791-470-7 (PDF)
ISBN: 978-1-80791-472-1 (Epub)

Typeset by TNQ Tech

CONTENTS

About the Authors ..... *vii*

Preface ..... *ix*

1. Introduction to Psychometric Concepts *1*
2. Item Analysis *21*
3. Reliability *41*
4. Generalizability Theory *71*
5. Validity *97*
6. Issues in Scoring *147*
7. Item Response Theory (IRT) *165*
8. Equating *255*
9. Differential Item Functioning *283*

References *331*

This page intentionally left blank

ABOUT THE AUTHORS

Holmes Finch is the George and Frances Ball Distinguished Professor of Educational Psychology, and Professor of statistics and psychometrics at Ball State University. His research interests include psychometrics, measurement, latent variable modeling, and nonparametrics.

Brian F. French is a Regents Professor of educational measurement, research methods, and statistics at Washington State University. His research sits at the intersection of applied and methodological issues in educational and psychological measurement.

Jason C. Immekus is Professor of Educational Measurement and Statistics at the University of Louisville. His research interests include latent variable modeling, psychometrics, and instrument development.

Maria Dolores Hidalgo Montesinos is Professor of Methodology of Behavioral Sciences at the University of Murcia (Spain) with over 34 years of experience in teaching and research in Psychometrics and Behavioral Science Methodology. Her work focuses on differential item functioning (DIF), item response theory (IRT), cross-cultural test adaptation, and the development of psychological tests and questionnaires.

This page intentionally left blank

PREFACE

This book was developed to provide a "one-stop" resource for many common analyses that an applied researcher might complete when working with various instruments to measure educational and psychological traits. We have developed examples, collected our favorite examples and resources, and provided explanations of analyses in one easily digestible text. Many of the analyses presented assist in providing the recommended evidence to support the inferences drawn from scores from such instruments. That is, the results from applying these techniques assist providing score reliability and validity evidence.

Through our years as graduate students and the first segment of our academic lives, we have explored the use of various programs for scale development and the study of the psychometric properties of the scores to provide such evidence. We have had the experience, as I am sure many of you have had, of turning to multiple books for instructions and examples to complete analyses in the scale development and validation process. For those readers just beginning on the psychometric ride, you too will soon see experience this. By no means will this book stop the need for multiple sources, in fact, that is always encouraged. However, this book should allow the reader to use this as a main guide and supplement to experience analyses described in major text books. Our examples are intended to be clear and concise with R syntax that can be easily be adapted to fit many situations, as the reader learns and uses various techniques.

The anticipated audience for this book includes researchers, practitioners, and graduate students searching for a guide to perform common psychometric analyses on various assessments. We assume a basic level of statistical knowledge but review concepts throughout. We envision that this text will (a) patiently wait on some office selves begging to be handed to a student as a resource, (b) have a permanent home on desks where it continually rises to the top of the stacks for daily use of the applied researcher,

(c) be happily carried in bags to and from work and class by the graduate student learning techniques, (d) be listed proudly as a reference text on syllabi, and finally (e) as an occasional drink coaster while deep thoughts are pondered about how to solve measurement problems. We hope that through such uses, particularly the latter, that we have provided some insight and assistance to the user in appropriately applying the techniques and concepts discussed.

We cover major topics such as item analysis, score reliability and validity, generalizability theory, differential item functioning, equating, and so on. Under each topic we present foundational ideas and give examples of how to apply these ideas immediately in ones work. Chapter 7, for instance, contains information on differential item functioning (DIF). We discuss DIF, its importance in the score validation process, and provide three techniques using R to detect DIF, including how to handle clustered data in such analyses. The caveat is we do not provide a detailed discussion of each topic but rather the essence of each topic and several resources for additional reading. Again, we remind you that this is just one resource to promote your psychometric knowledge and practice.

We do assume the user has some basic knowledge and skill level in operating R. However, we do provide some refreshers (e.g., reading in various files, formatting) in Chapter 1 for those who have not had the luxury of using R recently. We do attempt to present material in a very understandable language avoiding or explaining jargon as we go. You may find the occasional joke thrown in from time to time to spice it up. But remember we are researchers, not comedians, even though students and family seem to laugh often at us for the things we think about!

We do ask that if you have feedback, efficiency techniques, improvements, or just plain find mistakes to please notify us. We welcome user feedback and will incorporate this into a revision, if demanded by the reader!

So with that, let us get started on our R adventure in applying psychometric techniques. In the words of Dr. Seuss "Take it slowly. This book is dangerous."

Enjoy,
Holmes Finch, Brian F. French, Jason C. Immekus,
Maria Dolores Hidalgo Montesinos

CHAPTER 1

INTRODUCTION TO PSYCHOMETRIC CONCEPTS

Measurement Basics

Measurement is a mainstay of educational and psychological practice and research. For instance, educators use academic achievement test scores to measure student and school performance, psychologists use scores from self-reported instruments to measure client mood, and colleges and universities use scholastic aptitude test scores to guide admissions decisions. In each instance, obtained scores play a critical role in decision making about individuals and groups. Therefore, the functioning of an instrument must be well understood and carefully studied to ensure that it provides relevant and meaningful information. Over the last roughly 100 years, a subspecialty combining statistics and educational psychology has developed to study such measures. This field, known as psychometrics, focuses on the development, and vetting of educational and psychological assessments using a range of statistical procedures. Together, empirical evidence obtained from psychometric investigations of item- and test-level scores offer vital information to guide score interpretation and use for designation purposes (e.g., diagnosis, placement). In this book, we cover a variety of procedures to investigate the psychometric properties, namely reliability and validity, of obtained scores to substantiate score use and/or scale development. Specifically, we focus on the R statistical software to investigate test score reliability and validity (https://r-project.org).

This book aims to promote your understanding and ability to implement the procedures found in subsequent chapters to analyze item- and test-level

Applied Psychometrics Using R Software, pages 1–19
© 2027 by Emerald Publishing Limited
All rights of reproduction in any form reserved.
doi:10.1108/978-1-80791-470-720261001

data obtained from diverse instruments (e.g., self-report survey) to guide your decision-making (e.g., scale development). Specifically, the book begins with introducing diverse item types, including dichotomous responses inclusive of two possible response options (e.g., true/false, correct/incorrect) and polytomous items, which include three or more response options (e.g., Likert scale items). At the scale level, we present methods for examining the psychometric properties of scores derived at the test level, including reliability (e.g., internal consistency) and validity (e.g., relationship to external variables). This includes the steps to conduct an item analysis to gather information on the functioning of items comprising an instrument, including item difficulty and discrimination. More specifically, on an academic achievement test, item difficulty represents how easy or difficult an item was for a group of students, whereas for a self-report measure of well-being item difficulty refers to how likely an individual (e.g., patient) is to endorse the item along the response continuum (e.g., *Strongly Disagree* to *Strongly Agree*). Correspondingly, item discrimination refers to how well an item differentiates among examinees with different standings on the measured trait (e.g., achievement motivation, depression). Throughout this book, we will refer to the individuals completing the items as examinees. Similarly, the terms instruments and tests may be used interchangeably since they represent approaches to gather data to determine individual and group standing on the measured trait. It is important to note that virtually all of the topics in this text are equally applicable to traditional tests of achievement or aptitude, as well as to affective and noncognitive assessments (e.g., self-efficacy, well-being). Finally, throughout the text we will discuss the notion of the latent trait being measured, which refers to the attribute we seek to measure, including intelligence, depression, or political outlook. The score obtained from the instrument will typically serve as the manifest indication of this unobserved, or latent, variable. Prior to getting into the nuts and bolts of how to analyze data using R, let us first discuss the two primary paradigms that underlie nearly all approaches to examine the psychometric properties of obtained scores, specifically: Classical Test Theory (CTT) and item response theory (IRT).

Classical Test Theory

Historically, CTT has served as the basis for much of what we think of as psychometrics and measurement. Specifically, over the last 100 years, CTT has framed the notion of test score reliability and, correspondingly, much of validity. In addition, although the form of the CTT model differs substantially from that of the later developed IRT model, subsequently discussed, they share many of the same basic concepts (e.g., item difficulty and discrimination). In general, CTT is simply a way to link an observed

score derived from an instrument to the unobserved attribute (e.g., motivation, depression) that we seek to measure. Thus, for example, if we administer a 5th grade class an interim math assessment, we probably want to use scores to tell us something about curriculum effectiveness and students' attainment of taught content. In an ideal world, we would want to be able to directly assess students' mathematics knowledge, but for reasons that will soon become clear, this is not possible. However, if our test is well designed and psychometrically sound, we can infer that obtained scores represent reasonably accurate and reliable estimates of mathematics achievement. In this section, we will discuss the ideas underlying CTT, and their implications for educational and psychological measurement.

In CTT, the relationship between an observed variable X and the underlying trait T is represented as $X = T + E$, where X is an individual's observed score on the measure, T is the individual's true score on the measured trait (e.g., depression, motivation), and E is random error. Put into words, this equation states that the observed score an individual receives on a test is a function of their true knowledge of the subject (assuming we're discussing some type of cognitive or achievement measure) and a set of other factors that are random in nature. Broadly, we can think of T as a stable characteristic inherent to the individual that would remain unchanged over repeated administrations of the instrument, if that could be done so that after each test examinees forgot that they had taken it (Haertel, 2006). Thus, in theory, if we were able to assess an individual over an infinite number of occasions, T would equal their average score across occasions. Correspondingly, error is generally conceptualized as the unstable factors that influence examinee performance on the measure (e.g., contextual distractions). One way to consider error is by classifying it into 4 distinct types, including: (1) natural variation attributed to the examinee on the day of assessment (e.g., fatigue, hunger, mood), (2) environmental factors present during test administration (e.g., room temperature, ambient noise), (3) scoring variation (e.g., ratings by evaluators), and (4) test items (Feldt & Brennan, 1989).

The random nature of error leads to a number of interesting properties of the CTT model. First of all, if a particular examinee could be given the test repeatedly over a very large number of times, and each time forget that (s)he had taken it, the mean of the errors across those test administrations would be 0 (i.e., the population mean, $\mu_E = 0$). Furthermore, the random nature of error leads to the underlying assumption in CTT that E is uncorrelated with T. In other words, if we had a group of students completing a test comprised of a large number of times, and calculated Pearson's r between the true score (T) and error (E), it would equal 0 (i.e., $r_{T,E} = 0$). In addition, if we had multiple forms of the same exam, the errors across those forms would also be uncorrelated, again because their errors are random, or $r_{E1,E2} = 0$.

While these results are interesting in and of themselves, they lead to a relationship that is key in CTT. Given that we know T and E are uncorrelated, we also know that the covariance between them (COV) is also 0. Therefore, we can rewrite the composite variance of X as

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{1.1}$$

This relationship is central to the idea of test reliability, which we discuss in some detail in Chapter 3. For the moment, we can simply define the concept of reliability as the ratio of variance in T to the variance in X, or

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \tag{1.2}$$

In Chapter 3, we detail how to estimate reliability for a scale using scale data.

Often in practice, however, test users are interested in learning something about the likelihood of an obtained score representing an examinee's true score on the measured trait. Though T can never be known, because it is unobservable, the basic elements of CTT provide an approach to construct a confidence interval within which we have a certain level of confidence (e.g. 95%) that our obtained score X represents T. To construct confidence intervals to generate statements in which we feel we may capture T at designated levels (65%, 95%, or 99%), it is critical to understand the standard error of measurement (*SEM;* Harvill, 1991). Theoretically, if we could give the same individual a measure many times, and each time they would forget they had taken the measure, we would obtain a distribution for X. With such a distribution, we could then calculate measures of central tendency (e.g., mean) and variability (e.g., range). Of relevance to understanding SEM is the standard deviation (σ) of the distribution of X. Specifically, if a sample of examinees (e.g., 100) completed an exam 1,000 times we would have a distribution of scores (X) for each examinee. Therefore, we could calculate the standard deviation for each set of examinee scores, which reflects the variability in examinees' scores. Given that we assume T is stable for an individual, these standard deviations would actually reflect the error variation for each individual. If we were to average these standard deviations across all of the individual examinees in a given sample, we would obtain the *SEM*. In other words,

$$SEM = \frac{\sum_{i=1}^{N} \sigma_{Ei}^2}{N} \tag{1.3}$$

where

N = our sample size.

Of course, in practice we will not be able to have examinees take a test multiple times while forgetting their previous completion, so consequently we will not have access to σ^2_{Ei}. However, using a bit of algebra, it is possible to obtain a method for estimating *SEM* that is based on easily obtained statistics such as a reliability estimate (ρ_{XX}) and the standard deviation of the observed test scores (σ_X). In this formulation:

$$SEM = \sigma_X\sqrt{1-\rho_{XX}} \tag{1.4}$$

As we will see in Chapter 3, the *SEM* can be used to construct a confidence interval around T for examinees in the sample.

Prior to completing our discussion of CTT, it is important to consider the issue of parallel forms of a measure. In specific instances, parallel forms of the same instrument are needed when we may seek to administer an instrument repeatedly within a short duration (e.g., weeks), such as when a test's administration is unexpectedly discontinued (e.g., examinee fatigue, noise) and it needs to be readministered. In CTT, the idea of parallel forms is important particularly to estimating score reliability. Multiple forms of an instrument are said to be strictly parallel when they are developed using identical test specifications, produce the same distributions of X (i.e., same mean and standard deviation), have equal covariances between any pairs of forms, and covary equally with other measures. Based on these premises, examinees will have the same T across forms. Finally, given the random nature of E, the covariance between error terms for two strictly parallel forms will be equal to 0. In Chapter 3, we revisit the concept of parallel forms when discussing the estimation of scale reliability.

Item Response Theory

IRT represents a collection of statistical models to characterize how examinees answer items comprising an instrument. Thus, it offers another theoretical framework to considering the psychometric properties of diverse instruments. Whereas the focus of CTT is typically, although not exclusively, considered at the total test score level with issues such as reliability and validity, IRT focuses its attention on the psychometric properties of scores at the item level. There are a variety of IRT models to examine the likelihood of a particular item response and, collectively, they all focus on the relationship among characteristics of items, examinees, and the probability of a given response, such as obtaining a correct response on a mathematics test or selecting *Agree* on a Likert scale item. As we will see, IRT models are

available for a range of item types, including dichotomous and polytomous item responses. In this section, we begin with introducing models for dichotomous data (e.g., correct/incorrect; yes/no) and, subsequently, models designed for polytomous data commonly encountered with affective rating scales comprised of items using graded response options.

Dichotomous Items

Probably the single most common family of IRT models are those within the logistic framework. For dichotomous items, there are three models that are commonly used and each differs from the others in terms of how much information they contain about the items. The simplest model is the 1-parameter logistic (1PL) model and it serves as our general (and hopefully gentle) introduction to IRT. The 1PL model can be expressed mathematically as:

$$P\left(x_j = 1 \mid \theta_i, a, b_j\right) = \frac{e^{a\left(\theta_i - b_j\right)}}{1 + e^{a\left(\theta_i - b_j\right)}} \tag{1.5}$$

where x_j is the binary response to item j (e.g., 0 or 1), in which we can code a correct response as 1 and incorrect as 0 if we consider a measure of academic achievement. The variable θ_i is the value of the latent trait being measured by the test (e.g., reading aptitude) for given examinee i. The 1PL model also contains two item parameter values, a_j, item discrimination, and b_j, item difficulty. For this model it is assumed that a_j is constant across items (i.e., we assume all items are equally discriminating), while b_j varies across items. In terms of their meaning, we can view b_j as an indicator of how likely an individual with low proficiency is to answer the item correctly, when discussing proficiency exams. It is important to note that item difficulty and examinee level on the latent trait are on the same scale, which is centered at 0 and theoretically ranges from -∞ to ∞, though in practice it typically lies between −3 and 3 (i.e., z-score scale; de Ayala, 2009). An item with a lower difficulty is relatively easier than those with higher difficulty values, such that examinees with lower levels of θ_i are more likely to answer it correctly. Item discrimination (a) therefore, reflects how well an item differentiates among those with greater or lesser amounts of θ_i, with larger values indicating greater such discriminatory power.

As previously indicated, the 1PL model is based on the assumption that all items have equal discriminatory power. This assumption differentiates it from the subsequent IRT models we present. Specifically, in some instances, test users want to assume that the value of $a_j = 1$, thus defining a special case of the 1PL known as the Rasch model. Therefore, we can think of the Rasch model as a special case of the 1PL in which each item has a discrimination value of 1. However, it should be noted that in the broader measurement

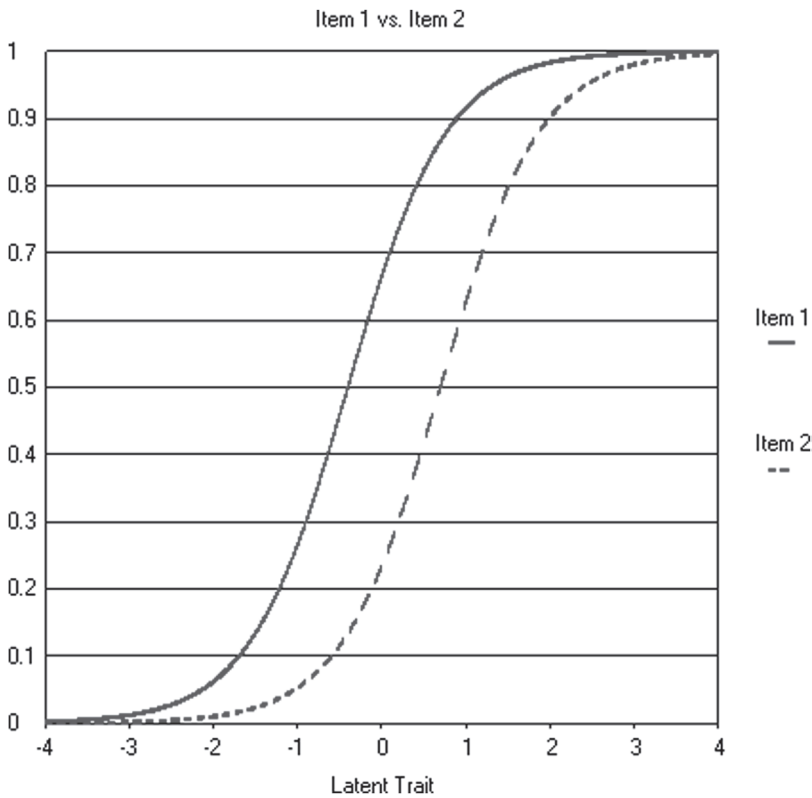
community the two models carry with them very different implications for practice. While will not discuss the issues surrounding the Rasch model further in this book, we encourage interested readers to investigate them in such resources as Embretson and Reise (2000) and de Ayala (2009), among others.

The item characteristic curve (ICC), also known as the item response function, is a common tool used to visualize the relationship between an examinee's latent ability level (θ) and the probability of a particular item response. Thus, the ICC relates the latent trait being measured (on the X axis), with the probability of a correct response (in the case of dichotomous items) on the Y axis. As an example, consider two items based on the 1PL model in which $b_1 = -0.4$, $b_2 = 0.7$ and $a = 1.2$. Fig. 1 below illustrates the ICCs for these two items.

Within Fig. 1, we can see that while the shape of the items (slopes) between Items 1 and 2 are the same, Item 2 is shifted to the right of Item 1, because

Fig. 1.

Item Characteristic Curve for 1PL Model.



it has a higher difficulty parameter value and, thus, is more difficult to obtain a correct response. As such, we use the ICC to determine the probability of a correct response for an individual with a given value of θ by drawing a straight line up from the *X* axis until it reaches the curve, at which point we would draw a second line from that point on the curve to the *Y* axis to obtain the probability of a correct item response. As an example, an examinee with an ability estimate of −1 (or, 1 standard deviation below the mean, 0) has a probability of approximately 0.30 of a correct response on Item 1 (score of 1), whereas on Item 2 the probability of a correct response is approximately 0.05.

When we cannot assume that item discrimination values are equal across the items, we can use the 2-parameter logistic (2PL) model, which has very similar form to the 1PL:

$$P\left(x_j = 1 \mid \theta_i, a_j, b_j\right) = \frac{e^{a_j\left(\theta_i - b_j\right)}}{1 + e^{a_j\left(\theta_i - b_j\right)}} \tag{1.6}$$

The difference is that now a_j is specific to item *j*. By allowing the discrimination parameter to vary across items, it is possible to identify those items that are best able to differentiate among individuals based on the level of the latent trait being measured, with higher values indicating an item that is more discriminating. The ICCs for a pair of items modeled using the 2PL model with item parameters $a_1 = 1.2, b_1 = -0.4, a_2 = 0.8,$ and $b_2 = 0.7$, respectively, appear in Fig. 2. As with the 1PL model, Item 2 is more difficulty as it is shifted to the right of Item 1. However, as shown, Item 1 has higher discrimination as its ICC is steeper compared to Item 2 which is slightly flatter.

A third variation on the logistic framework for dichotomous items is the 3-parameter logistic model (3PL), which incorporates the probability that an examinee will provide a correct item response simply due to chance, perhaps by guessing. The 3PL model is expressed as

$$P\left(x_j = 1 \mid \theta_i, a_j, b_j\right) = c_j + \left(1 - c_j\right)\frac{e^{a_j\left(\theta_i - b_j\right)}}{1 + e^{a_j\left(\theta_i - b_j\right)}} \tag{1.7}$$

where c_j is the pseudo-chance (or pseudo-guessing) parameter, representing the probability of a correct response for an examinee whose value of θ_i approaches -∞. As an example, consider a multiple-choice test item with 5 response options to answer an item, such as: A, B, C, D, or E. An individual with an extremely low value of θ_i (e.g., −2) would also have an extremely low probability of answering the item correctly, based on their knowledge of the subject being tested. However, for a multiple-choice type