A Course in Item Response Theory and Modeling with Stata

Tenko Raykov
Michigan State University

George A. Marcoulides
University of California, Santa Barbara

A Stata Press Publication
StataCorp LLC
College Station, Texas

Copyright © 2018 StataCorp LLC
All rights reserved. First edition 2018

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-266-4
Print ISBN-13: 978-1-59718-266-9
ePub ISBN-10: 1-59718-267-2
ePub ISBN-13: 978-1-59718-267-6
Mobi ISBN-10: 1-59718-268-0
Mobi ISBN-13: 978-1-59718-268-3

Library of Congress Control Number: 2017957532

No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without the prior written permission of StataCorp LLC.

Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LLC.

Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations.

NetCourseNow is a trademark of StataCorp LLC.

LaTeX 2ε is a trademark of the American Mathematical Society.

Contents

(Pages omitted)

Preface

More than half a century ago, a far-reaching revolution started in behavioral, educational, and social measurement, which to date has also had an enormous impact on a host of other disciplines ranging from biomedicine to marketing. At that time, item response theory (IRT) began finding its way into these sciences. In many respects, IRT quickly showed important benefits relative to the then conventional approach for developing measuring instruments that was based on "classical" procedures.

Since the 1950s and the influential early work by F. Lord in IRT (for example, Lord [1952, 1953]), more than 60 years have passed that have been filled with major methodological advances in this field and more generally in behavioral and social measurement. The intervening decades have also witnessed an explosion of interest in IRT and item response modeling (IRM) across those disciplines as well as the clinical, biomedical, marketing, business, communication, and cognate sciences. These developments are also a convincing testament to the rich opportunities that this measurement approach offers to empirical scholars interested in assessing various latent constructs, traits, abilities, dimensions, or variables, as well as their interrelationships. The latent variables are only indirectly measurable, however, through their presumed manifestations in observed behavior. This is in particular possible via use of multiple indicators or multi-item measuring instruments, which have become highly popular in the behavioral and social sciences and well beyond them.

This book has been conceptualized mainly as an introductory to intermediate level discussion of IRT and IRM. To aid in the presentation, the book uses the software package Stata. This package offers, in addition to its recently developed IRT command, many and decisive benefits of general purpose statistical analysis and modeling software. After discussing fundamental concepts and relationships of special relevance to IRT, its applications in practical settings with Stata are illustrated using examples from the educational, behavioral, and social sciences. These examples can be readily "translated", however, to similar utilizations of IRM also in the clinical, biomedical, business, marketing, and related disciplines.

We find that several features set our book apart from others currently available in the IRT field. One is that unlike a substantial number of treatments of IRT (in particular older ones), we capitalize on the diverse connections of this field to the comprehensive methodology of latent variable modeling as well as related applied statistics frameworks. In many aspects, it would be fair to view this book as predominantly handling IRT and IRM, somewhat informally stated, as part of the latent variable modeling methodology. In particular, the discussion throughout the book benefits as often as possible from the

conceptual relationships between IRT and factor analysis, specifically, nonlinear factor analysis. Relatedly, whenever applicable, the important links between IRM and other statistical modeling approaches are also pointed out, such as the generalized linear model and especially logistic regression. Another distinguishing feature of the book is that it is free of misconceptions about and incorrect treatments of classical test theory (Zimmerman 1975). Regrettably, they can still be found in some measurement literature and inhibit significantly in our opinion progress in social and behavioral measurement. In addition, these misconceptions contribute to a compartmentalization approach that seems to have been at times followed especially when disseminating or teaching IRT in circles with limited or no prior familiarity with it. That approach and resulting restrictive focus of interest is in our view highly undesirable. The reason is that such an approach has the potential of creating long-term disservice to the cause of behavioral and social measurement. In this connection, we would also like to point out that unlike many previous treatments, this book presents its discussion and developments without any juxtaposition of IRT to classical test theory. This is because IRT does not need this kind of "comparison" and related misconceptions to convince scholars of what it can deliver under its assumptions (see also Raykov and Marcoulides [2016b]). A third characteristic of the book is that it demonstrates the straightforward, user-friendly, and highly effective Stata applications for IRT modeling. We hope to gain in this way many new enthusiasts for this methodological field as well as IRT software across these and related disciplines. Last but not least, our book aims to provide a coherent discussion of IRT and IRM independently of software. The goal is thereby to highlight as often as possible and in as much detail as deemed necessary important concepts and relationships in IRT before moving on to its applications. This was necessary because in our experience, many individuals seem to find some features of this modeling approach more difficult to deal with and use to their advantage than what may be seen as "conventional" applied statistical concepts and relationships. These features include in particular the inherent nonlinearity in studied item-trait relationships as well as produced estimates (predictions) of individual trait levels and measures of uncertainty associated with them. That difficulty in appreciating characteristic properties of IRT may have arguably resulted from insufficient discussion and clarification of them in some alternative accounts or presentations.

This book could be considered aimed mainly at students and researchers with limited or no prior exposure to IRT. However, we are confident that it will also be of interest to more advanced students and scientists who are already familiar with IRT, in particular owing to the above mentioned features in which the book does not overlap with the majority of others available in this field. In addition, a main goal was to enable readers to pursue subsequently more advanced studies of this comprehensive and complex methodological field and its applications in empirical research, as well as to follow more technically oriented literature on IRT and IRM. Relatedly, even though the book uses primarily examples stemming from the educational and behavioral sciences, their treatment, as well as more generally of this measurement field, allows essentially straightforward applications of the used methods and procedures also in other social science settings. These include the clinical, nursing, psychiatry, biomedicine, criminology,

organizational, marketing, and business disciplines (for example, Raykov and Calantone [2014]).

Our book has been influenced substantially by deeply enriching interactions with a number of colleagues over the past years. Special thanks are due to K. L. MacDonald and R. Raciborski for their many instructive inputs on Stata uses and applications in relation to examples used in the book, as well as on IRT and its empirical utilizations in more general terms. The importance of the contributions also of Y. Marchenko and C. Huber cannot be overstated, who provided instrumental support during our work on the book. We are especially indebted to C. Huber for helpful comments and criticism on an earlier version, which contributed markedly to its improvement. His assistance during the book-production phase was similarly invaluable, as was that of the book editor and the production assistant. We also wish to express our particular gratitude to M. D. Reckase, B. O. Muthén, D. M. Dimitrov, M. Edwards, C. Lewis, R. Steyer, S. Rabe-Hesketh, A. Skrondal, and A. Maydeu-Olivares for valuable discussions on IRT and IRM and related applied statistics and measurement approaches. We are similarly thankful to C. Falk, R. J. Wirth, N. Waller, R. Bowles, I. Moustaki, R. D. Bock, S. H. C. duToit, G. T. M. Hult, and J. Jackson for insightful discussions on IRT applications and software. We are also grateful to a number of our students in the courses we taught over the last few years who offered very useful feedback on the lecture notes we first developed for them, from which this book emerged. Last but not least, we are more than indebted to our families for their continued support in lots of ways that cannot be counted. The first author is indebted to Albena and Anna; the second author is indebted to Laura and Katerina.

Tenko Raykov and George A. Marcoulides

(Pages omitted)

1 What is item response theory and item response modeling?

1.1 A definition and a fundamental concept of item response theory and item response modeling

Item response theory (IRT) is an applied statistical and measurement discipline that is concerned with probabilistic functions describing i) the interaction of studied persons and the elements of a measuring instrument or item set of concern, such as items, questions, tasks, testlets, subtests, and subscales (generically referred to henceforth as "items"); and ii) the information contained in the data, which are obtained using the instrument, with respect to its items and overall functioning as well as the examined persons (Reckase 2009).

A fundamental concept in IRT is the relationship between i) the trait, construct, ability, or latent dimension (continuum) being evaluated with the instrument, the dimension being typically denoted θ and often assumed unidimensional but in general may consist of two or more components (see below); and ii) the probability of "correct" response on a given item for a random subject with a trait or an ability level, θ, that is designated as $P(\theta)$.[1] A function of θ, which describes this probability $P(\theta)$ for an item, is called an item characteristic curve (ICC). (Throughout this chapter, we assume that θ is unidimensional unless otherwise indicated.) Owing to its special relevance to IRT and item response modeling (IRM), the ICC can be viewed as one of its main concepts. Other frequent references to it are item response curve, item characteristic function, item response function, or item trace curve. It is important to stress that while being defined as a probability, the ICC is not assumed to be a "static" concept but is rather a function of the (presumed) underlying latent dimension θ. This functional relationship between the probability of a particular response ("correct" response) on a given item and the studied trait, construct, or ability, θ, can be viewed as a characteristic

1. As is common in the IRT literature, the notation θ is used throughout this book to denote i) the studied latent trait, ability, construct, continuum, or, in general, latent dimension (or dimensions); ii) an individual value or point on the last (for example, a subject's latent trait or construct score or ability level that is of interest to evaluate); and iii) the horizontal axis of figures of item characteristic curves (ICCs) that imply the consideration or assumption of θ as a single latent continuum or "scale", which is at times also referred to as "θ-scale". While this might be viewed potentially as mixing or even as overusing the symbol θ, strictly speaking, we follow this standard notation that has been widely used in the literature over the past several decades. The specific meaning of its usage, in a sense mentioned in points i) through iii) in this note, is determined by context in the pertinent discussions in the following chapters.

element of IRT and IRM. A prototypical ICC, for the case of a single latent dimension or continuum θ that underlies subject performance on a given set of items or measuring instrument (and a binary scored item), is presented in figure 1.1, which also emphasizes the extended S-shape of this curve.

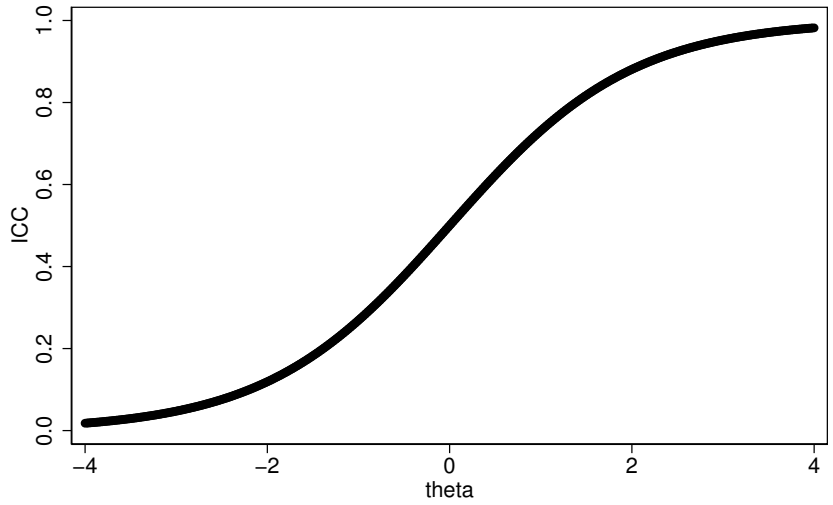

Figure 1.1. A prototypical item characteristic curve

From figure 1.1, we can see that as one moves from left to right on the horizontal axis representing the studied latent dimension θ (referred to as "theta" on the figure), the ICC is fairly low initially, then enters a region of notable increase in the central part of the curve; toward the right end of the presented range of θ, the curve "stabilizes" at a fairly high level, approaching 1 (see chapter 2 for further discussion on the ICC concept). As will be explicated later, different items in a given measuring instrument or item set of concern differ in general in the steepness of the curve increase in that central part of the ICC. The extent of this increase, or curve tilt, is captured by a particular parameter of interest in IRT and its applications in instrument construction and development, which is referred to as item discrimination parameter. (In chapter 5, we discuss in detail this and other item parameters.) Although the majority of applications of IRT are currently based on models that assume the same functional class for the ICCs of all items involved in a studied set or instrument (for example, when all items are binary or binary scored), one can also use IRT with so-called hybrid models. These models allow subsets of items to follow different functional classes for their ICCs, for example, when some items are binary whereas others are ordinal and with more than two available response options (see chapter 11).

While individual items are a special focus of IRT and IRM, measuring instruments consisting of multiple items are also of particular interest. Such instruments—for instance scales, tests, test batteries, surveys, questionnaires, self-reports, inventories, subscales, or testlets—are highly popular in the behavioral, educational, and social sciences (for example, Raykov and Marcoulides [2011]). Their popularity in these and cognate disciplines is to a large degree due to their being composed of multiple components, which provide converging pieces of information about underlying traits, abilities, and attitudes, or in general latent dimensions that are often referred to as "constructs". These constructs and their relationships with one another and with other variables are of main interest in those and related sciences. The reason is that entire theories in them are advanced and developed in terms of such indirectly observable, latent, or hidden variables. This is because the latter typically reflect substantively important theoretical concepts of special concern in these and cognate disciplines. The latent variables can be defined as random variables that presumably possess individual realizations in all subjects in a studied population (or a sample from it), while no observations are available on their realizations (for example, Bollen [1989]). These variables are unobserved, however, because they cannot be directly measured, assessed, or evaluated. They are assumed to be continuous throughout this book, and information about them is collected in their manifestations, proxies, or indicators in observed behavior. (See, for instance, Raykov, Marcoulides, and Chang [2016] and references therein for alternative settings with discrete latent variables.) As such latent variable manifestations, one can usually consider the responses obtained from the studied subjects on the components or elements of instruments used to evaluate the unobserved constructs. Thereby, in the role of instrument components, one typically uses appropriate items, such as questions, tasks, or problems to solve (for example, McDonald [1999]).

The present book, as alluded to earlier, deals with a particular approach to the study of the interaction of persons (respondents, examinees, patients, etc.) with measuring instruments and especially with their elements or components (items). The aim thereby is to optimally use the information about persons and items that is contained in the subject responses to the items. We will be specifically concerned with these responses on the items as well as the studied persons' performance on the considered instruments.

As indicated above, a major focus in IRT is on the relationship between i) the probability P of a particular type of response (such as "correct", "true", "present", endorsed, "success", "agreed", "applicable", etc.) on any given item; and ii) the underlying presumed latent dimension (or dimensions) of interest to evaluate, such as ability, proficiency, trait, construct, or attitude (latent continuum), typically denoted θ in IRT contexts (see also footnote 1). That is, throughout this book, we will be especially interested in the above function $P(\theta)$ describing this relationship—called ICC as mentioned earlier—for each of the items in a measuring instrument or item set under consideration. With this in mind, one could define IRT in simple terms as a methodology dealing with modeling the function $P(\theta)$, that is, expressing in quantitative terms the relationship between θ and the above probability as a function of θ. This probability function includes specific characteristics (parameters) of the used items and of the studied subjects, with the items usually representing a measuring instrument of interest. In fact, most

contemporary IRT applications can be seen as essentially concerned with the following activities (no ranking is implied in terms of their relevance):

a) postulating models about this relationship, that is, for $P(\theta)$, that involve unknown parameters associated with the items of the instrument;

b) estimating these parameters using an available dataset obtained with the instrument;

c) evaluating the (relative) fit of the models used; and,

d) based on the results of the activities in a) through c), estimating (predicting) individual subject values for θ using plausible (selected, preferred) models, with the values being positioned on the same "scale" or underlying latent dimension or continuum as are particular item parameters (when θ is unidimensional; see below).

An especially important and useful feature of IRT is that at the end of its application, using a plausible model for an available dataset from a given item set or instrument one obtains the following two sets of quantities that are commensurate, that is, located on the same continuum (when unidimensional) (for example, van der Linden [2016b]):

i) a set of (estimated) quantities or parameters characterizing the items, specifically, their difficulty parameters (see chapter 5 for a more precise definition); and

ii) a set of quantities or values (predictions, assigned values, or estimates—one per person in unidimensional IRT and more than one in multidimensional IRT) that characterize the extent to which each person possesses the trait or traits being evaluated with the instrument in question (see chapters 5, 6, and 12 for further details and examples).

Based on this discussion, we can observe the following important fact. For a given person and binary or binary scored item (measure), the function $P = P(\theta)$ is actually the mean of his or her response or observed or recorded score on the item. This observed score is in general a random variable, and we will denote it by Y throughout the book, also when it is nominal or ordinal with more than two possible responses (see Raykov and Marcoulides [2013] and also chapters 3 and 11). We emphasize that it is the latent dimension (or dimensions), θ, that is of actual interest to measure. However, as we indicated earlier, this is not possible to achieve in any way similar to how one can measure, say, length or weight. In particular, there is no "ruler" or (weight) "scale" that is available to accomplish this measurement. Instead, only the crude assessment of θ is feasible, namely, by using the above indicated measurement process using the presumed proxies, indicators, or manifestations of θ, that is, the items or instrument components. This process consists of administering the set of items or instrument used and recording subject responses to them. The process produces as a result the observed (recorded,

manifest) Y scores of the studied persons on the individual items. This complex set of what are actually indirect measurement activities with respect to θ that yield in the end the observed scores on the items is followed in IRM by suitable modeling and estimation procedures that are of main concern in the remainder of the book.

As far as the ICCs are concerned, we should stress that there are infinitely many possible choices of the function $P(\theta)$ for any given item (instrument component). However, as seen later in this book, only a few have obtained prominence and are used most of the time in current behavioral, educational, and social research. The specifics of these choices are attended to in later chapters.

A simple representation of the aims of IRT and IRM can be found in the following schematic, figure 1.2. This figure is used only for conceptual purposes here and is not meant to be a model or "causal" path diagram; that is, no causal implications are intended to be drawn from it ($k > 1$ is the number of items in a set, test, scale, or, more generally, measuring instrument of concern). The graphic representation in figure 1.2 makes use for the current aims of what has been often referred to as path diagram notation (for example, Raykov and Marcoulides [2006; 2011]). In this widely used notation, i) rectangles denote observed variables, or items (that is, the variables that we collect or record data on), while ii) ellipses symbolize the unobservable (latent) variable or variables that we are interested in making inferences about based on the obtained data on the items. As pointed out above, it is the latent variable, trait, or construct—in general, latent dimensions or continua—that the observed measures presumably contain information about. Hence, it is of interest to "extract" this information via appropriate use of optimal statistical methods, like those offered by IRT.

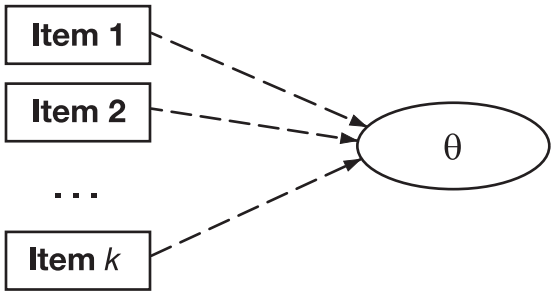

Figure 1.2. A conceptual representation of the aims of item response theory and item response modeling

In figure 1.2, on the left is what we observe, that is, have or collect or record data on, namely, the k items usually representing a measuring instrument of concern. On the right of this figure is what we want to make inferences about using those data—namely, the latent trait, construct, or ability, θ. The connections between the items and θ, represented informally by the one-way arrows in the middle of the figure, are facilitated by the assumed ICC for each item when θ is unidimensional (see chapter 12 for the multidimensional case). These ICCs are typically taken to be monotonically increasing

(continuous) functions of θ. (We emphasize that these item-trait relationships need not be assumed linear, with more details following below.)

This book mainly considers the underlying trait, ability, or construct as unidimensional, that is, as a single continuum [that is, with $\dim(\theta) = 1$, where $\dim(\cdot)$ denotes dimensionality or number of dimensions]. In addition, however, in a later chapter, we also provide an introduction to multidimensional IRT and IRM, where more than one trait or abilities are of concern, that is, $\dim(\theta) > 1$ holds (see chapter 12). The book will not deal with time-limited, timed, or speeded tests or behavior measuring procedures or with models for continuous responses. (See van der Linden [2016a] for discussions of such types of procedures.) In this respect, we observe that if a test is speeded (to some "significant" degree at least), then in general, it cannot be really considered unidimensional. The reason is that it will possibly be measuring the trait of initial interest as well as the ability to perform under a speeded condition (speed performance).

1.2 The factor analysis connection

Two applied statistics and measurement fields that are closely related to IRT and also of special relevance in the remainder of this book are classical test theory (CTT) and factor analysis (FA). We will be concerned in more detail with them in chapter 3, but it is useful here to highlight their links from a historical viewpoint to IRT and IRM. The basics of CTT and FA were laid in the beginning of the past century, starting perhaps with the far-reaching work by Spearman in two landmark papers published in the *American Journal of Psychology* more than 100 years ago (Spearman 1904b,a).

For a number of years, especially in the first half and middle of the 20th century, FA was focused on analysis of correlation matrices for observed variables measured on an interval scale (or treated otherwise as continuous). A main characteristic of those developments, which led to the classical (linear) FA model, was the assumed linear relationship between observed responses and underlying continuous latent variables. The latter variables are typically called factors in the FA context and are in general closely related to the latent dimensions θ in IRT (see chapters 3 and 4 for more detail). In the 1960s and 1970s, work began on extending the ideas of FA to the case of discrete observed variables and nonlinear trait-response relationships. This research evolved within a more general latent variable modeling (LVM) framework (for example, Muthén [2002]) and was led by K. G. Jöreskog and his collaborators, most notably D. Sörbom, A. Christofferson, and B. O. Muthén.

As a highlight of this early work, Muthén (1984) developed a comprehensive and widely applicable approach to FA-based modeling of discrete responses. This approach no longer assumed linear relationships between observed variables (indicators) and underlying factors. In this connection, it is also important to emphasize the contribution by McDonald (1967) that precipitated the extension of the FA framework to the case of nonlinear relationships between response and latent variables. His influential work is perhaps best understood in the context of the generalized linear model (GLIM) (see also

McDonald [1999]; compare Raykov and Marcoulides [2011]). The GLIM is also discussed in one of the following chapters and referred to on several occasions later in the book.

In a related development, Takane and de Leeuw (1987) showed that the single-factor model for discrete items (binary or binary scored indicators) is equivalent to the two-parameter normal ogive model. (See also Raykov and Marcoulides [2016b] for closely related results and implications for the relationship between CTT and IRT.) This research, which in our view unfortunately remained underappreciated throughout most of the ensuing decades, effectively demonstrated in tandem with McDonald (1967) the equivalence of a large class of IRT models to nonlinear FA models with discrete dependent variables. Much of the discussion in the rest of this book can in fact be seen as capitalizing on these earlier developments in the nonlinear FA and IRT fields, which share important commonalities and overlap more than may appear at first. (For a discussion on some of these commonalities, see, for instance, Kamata and Bauer [2008] and Raykov and Marcoulides [2016b].)

For the most part, the present book is based as mentioned on the important connections that exist between LVM and, in particular, FA on the one hand and IRT and IRM on the other hand. We find these links regrettably still not used or explicated in many IRT treatments. The unique benefits of these connections lie in the fact that they provide important insights into the deeper relationships between these two main applied statistics frameworks, allowing clearer understanding of each one of them. In this way, their links serve the aim of facilitating a more thorough understanding of IRT—as well as of LVM—that is void of potentially misleading and in our opinion outdated views of IRT (or of FA for that matter). It is those views, at times still disseminated among some circles, that we find to hamper progress in either of these highly popular and closely related applied statistical fields in contemporary behavioral and social science. These are views that this book is free of.

1.3 What this book is, and is not, about

This book will be concerned predominantly with unidimensional IRT and IRM and will in addition provide an introduction to multidimensional IRT (see chapter 12). Although the book deals chiefly with binary or binary scored items, it also discusses in considerable detail in a separate chapter polytomous items and IRT models for them (see chapter 11).

While being concerned with these theoretically and empirically important settings, as indicated earlier, the book will not be dealing with

- speeded, timed, or time-limited tests (see, for example, van der Linden [2016a], for a discussion of such tests);
- measuring instruments yielding continuous outcomes, that is, continuously distributed "items" (models for such instruments or components or items are referred to at times as "continuous response models"; for example, see Hambleton, Swaminathan, and Rogers [1991] and van der Linden [2016a] for references);

- settings with clustered (nested) data (for example, Hox [2010]);
- settings with considerable unobserved heterogeneity, so-called mixture IRT (for example, see Lubke and Muthén [2005]; see also Raykov, Marcoulides, and Chang [2016]);
- nonparametric IRM (for example, Sijtsma and Molenaar [2002]); and
- extended IRM with covariates (for example, Bock and Moustaki [2007]).

Based on the discussions provided in this book, it is also our goal to enable the readers to move on subsequently to more advanced treatments and uses of IRT and IRM, such as advanced multidimensional and polytomous IRT modeling, as well as IRM with clustered data (multilevel IRT), latent class-based (mixture) IRT, and IRT models with covariates.

1.4 Chapter conclusion

In broad terms, IRT can be seen as based on two postulates. One of them is that subject responses on a given set of items or measuring instrument can be explained by one or more latent traits, abilities, constructs, dimensions, continua, or factors. According to the second postulate, the relationships between the item responses and these traits can be described by appropriate probabilistic functions, which represent the ICCs in unidimensional IRT (compare Hambleton, Swaminathan, and Rogers [1991]; see, for example, figure 1.1 and chapters 2, 5, and 12). An examined person's responses on the items are assumed to depend thereby on i) the degree to which he or she possesses the studied traits and ii) one or more characteristics of each of the items. This relationship is described by specific probabilistic functions, as assumed within the IRT models, that are postulated at the item level. IRT and IRM can thus be viewed as an applied statistical discipline that deals with a family of models for the ICCs associated with given sets of items or measuring instruments administered to studied persons in the unidimensional trait or ability case. These types of models include at least one parameter for each item and, in the general case, at least one parameter associated with each subject. There is a strong connection between IRT and FA, especially nonlinear FA (McDonald 1967; Takane and de Leeuw 1987). Similarly, IRT is closely related to GLIM, in particular logistic regression, and to nonlinear regression (Cai and Thissen 2015). There are also strong connections between IRT and CTT (for example, Raykov and Marcoulides [2016b]). In fact, as will be illustrated in subsequent chapters, most popular IRT models can be viewed as nonlinear FA models that are empirically indistinguishable from appropriate CTT-based models in the single occasion assessment setting of relevance throughout this book (Raykov and Marcoulides 2011). These connections will be very beneficial for our discussions in the next few chapters as well as later in the book. The highlighting of the particularly useful IRT-FA-CTT-GLIM-logistic regression links, free of misconceptions about CTT and of its juxtaposition to IRT, is in fact a main distinguishing characteristic of this book.

(Pages omitted)

We recall from chapter 2 that the graphs of the logistic and normal ogive functions are indistinguishable for most practical purposes, after a minor rescaling of the horizontal axis for the former (with no substantive meaning in empirical settings in general). Thus, we can use figure 5.2 as a graphical representation of either of these functions for ICC illustration purposes in the rest of the book. (The units on the horizontal axis of the figure are not relevant for the present discussion.) In this figure, to prepare for the discussion in the remainder of the current and following chapters, we denote the horizontal axis by θ ("theta"), which is the underlying trait, construct, or ability of central interest in (unidimensional) IRT and its applications (see also chapter 1 and footnote 1 to it).

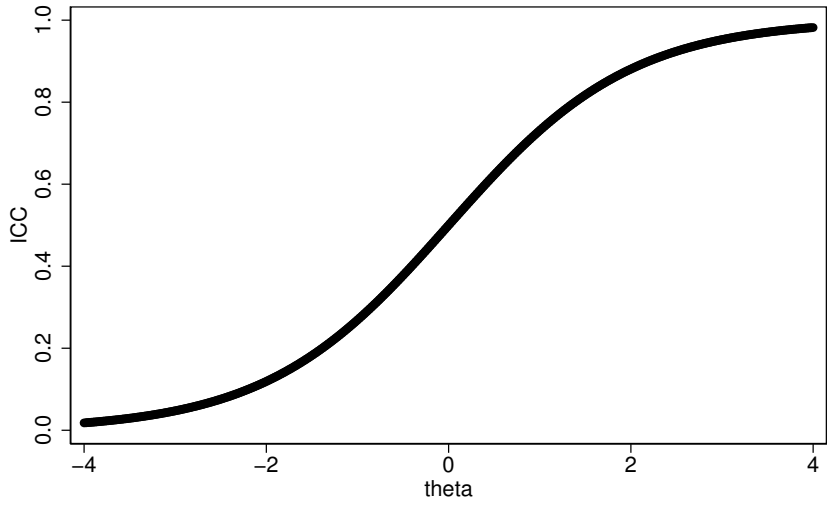

Figure 5.2. A typical (in shape) item characteristic curve in logistic item response theory models

We are now ready to discuss specific logistic IRT models that represent the overwhelming majority of IRT applications in contemporary behavioral, educational, and social research. These models are also used in a number of cognate disciplines ranging from biomedicine through marketing (for example, Raykov and Calantone [2014]).

5.5 The one- and two-parameter logistic models

When a logistic function is used as an ICC in a unidimensional IRT model, the height of the pertinent logistic curve at any given value of x, $\Lambda(x)$, obtains a special meaning, as it follows from (5.6) and is seen from figure 5.2. Specifically, this curve value $\Lambda(x)$ informs about the (assumed) proportion of persons in a population under investigation

at that trait or ability level who can answer "correctly" the corresponding item. That is, from (5.5) and (5.6), it follows that the associated probability of correct response on the item is

$$P(\theta) = \int_{-\infty}^{x} \psi(u)du = \Lambda(x) \tag{5.7}$$

In (5.7), $\Lambda(x)$ denotes as before the standard logistic cumulative distribution function, and $\psi(u)$ is its corresponding probability density function [compare Roussas [1997] and see (5.5)]. Using the inverse of the logistic function (see chapter 4), (5.7) is interpretable as saying the following:

$$x = \Lambda^{-1}\{P(\theta)\} \tag{5.8}$$

We obtained (5.8) by taking the inverse function of both sides of (5.7) (namely, its first and last part), recalling thereby that the inverse of the logistic function exists because of the latter being monotonically increasing (see also chapter 2). We also kept in mind that inverse functions applied successively simply annihilate or wipe out each other, as indicated earlier.

Equation (5.8) actually states that x is a function of θ, as we can see by looking at its right-hand side. Indeed, because the right-hand side of that equation is a function of θ, namely, $\Lambda^{-1}\{P(\theta)\}$, its left-hand side is also a function of θ, that is, x is a function of θ as well. Hence, we can rewrite (5.8) now as follows to emphasize this dependence (see in particular the last part of the next equation):

$$x = \Lambda^{-1}\{P(\theta)\} = x(\theta) \tag{5.9}$$

We can interpret (5.9) as demonstrating the following important fact when considering the logistic function giving rise to the ICC of a binary or binary scored item of concern. Specifically, to obtain the point x at which the value of the ICC equals a given probability, say, P, we need to take the inverse of the function representing the cumulative distribution function of the standard logistic distribution, $\Lambda(\cdot)$, at that value P. That inverse function depends on the studied latent dimension θ, as we have just seen, and this is why we denoted it $x(\theta)$ in (5.9).

With this discussion in mind, we can now obtain one of the most popular logistic IRT models by taking one more "small" step as we do next.

5.5.1 The two-parameter logistic model

A widely used model in the behavioral, educational, and social sciences when there is no guessing (or only minimal such that can be treated as negligible) on any item in a given set or measuring instrument is based on (5.9). (For instance, such are items involving free response or items that are administered after effective instruction in an assessment setting.) This widely used model assumes that for a given binary or binary

scored item, the dependence of x on the studied trait level θ as reflected in that equation is representable by the following linear relationship:

$$x = a(\theta - b) \tag{5.10}$$

In (5.10), a and b are parameters with special interpretation that are discussed in detail below. Because for a given item this model involves two parameters, it is referred to generically as a two-parameter model.

With (5.4) and (5.10) in mind, if we assume the logistic function as ICC for each item in a measuring instrument or item set of concern, from (5.7) follows that the formal representation of the ICC, that is, of the probability of "correct" response, is

$$P(\theta) = \Lambda(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}} = 1/\{1 + \exp(-x)\} \tag{5.11}$$

That is, owing to (5.10), from (5.11), we obtain

$$P(\theta) = \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} = \frac{1}{1 + e^{-a(\theta-b)}} = \frac{1}{1 + \exp\{-a(\theta - b)\}} \tag{5.12}$$

An IRT model that has as associated ICCs the function in (5.12) for each in a given set of binary or binary scored items, with in general different values for its parameters a and b, is called a two-parameter logistic (2PL) model.

We stress that (5.12) defines an item-specific model. This model, when considered in an empirical setting, is usually assumed for any item in an instrument or item set of interest. That is, if the model is posited for, say, the jth item, in a set consisting of k items ($k > 1$), to be more informative, one should attach the subindex j to both a and b in the right-hand side of (5.12). This leads to the following widely used form of the 2PL model (for the jth item),

$$P_j(\theta) = \frac{e^{a_j(\theta-b_j)}}{1 + e^{a_j(\theta-b_j)}} = \frac{1}{1 + e^{-a_j(\theta-b_j)}} = \frac{1}{1 + \exp\{-a_j(\theta - b_j)\}} \tag{5.13}$$

$(j = 1, \ldots, k)$.

We will discuss in detail the meaning of the a and b parameters in the next subsection. Before doing so, however, we observe an important fact that follows from (5.12) [see also (5.13)]. For a fixed value of the quantity b (and θ), an increase in the quantity a leads to an increase in the probability $P(\theta)$ (for "correct" response). Conversely, a decrease in a then brings about a decrease in that probability. Similarly, for a fixed value of a (and θ), an increase in the quantity b leads to a decrease in the probability $P(\theta)$. Alternatively, a decrease in b then yields an increase in that probability.

Keeping in mind the 2PL model, we now note a consequential link between IRT and IRM on the one hand and our discussion in the preceding chapter on the other. More concretely, through simple algebra on (5.13), we obtain

$$P_j(\theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}} = \frac{1}{1 + \exp\{-a_j(\theta - b_j)\}} = \frac{1}{1 + \exp\{-(c_j + a_j\theta)\}} \tag{5.14}$$

for the jth item, where $c_j = -a_j b_j (j = 1, \ldots, k)$. Comparing now each of the k equations in (5.14) with the corresponding equations (for the same items) in (4.12) in chapter 4 defining the multivariate logistic regression model with a single observed predictor, we can make the following observation. We can view the 2PL model, and by implication the Rasch model as a special case of it (see further details below), as a multivariate logistic regression model with i) a single unobserved predictor, namely, θ; ii) intercept $c_j = -a_j b_j$; and iii) slope a_j for the jth item in a set of items or measuring instrument of concern ($j = 1, \ldots, k$; compare Cai and Thissen [2015]). We will provide an alternative yet equivalent view of this relationship later in the chapter, after attending next to the particular meaning of the item parameters in this model.

5.5.2 Interpretation of the item parameters in the two-parameter logistic model

What do these parameters a_j and b_j actually mean in the 2PL model in relation to the jth item ($j = 1, \ldots, k$)? The parameter a_j can be shown to be directly proportional to the steepness of the ICC at its inflection point (for example, Reckase [2009]). In particular, the higher a_j, the steeper the slope of the ICC in its central part, and conversely (see figure 5.2). This inflection point, as mentioned in chapter 2, is located at that trait or ability level (point on the horizontal axis in an ICC graph) where the probability of correct response is 0.5. (See, for example, figure 2.2 in chapter 2 or figure 5.2 above. We mention in passing that this trait or ability level or point on the θ-scale is at times also referred to as the "midprobability" point for the considered item.) Looking at (5.12), we readily realize that this happens precisely where $\theta = b_j$ on the underlying continuum representing the studied trait or ability. (The reason is that only at this point, the numerator of the ratio in the right-hand side of this equation is equal to 1 and its denominator is equal to 2.) Thus, the meaning of the parameter b_j is as that position on the latent ability or construct scale (dimension or continuum), where $P_j(\theta) = 0.5$ holds for the probability of correct response on the jth item (see below for additional discussion).

(Pages omitted)

```
Contains data from http://www.stata-press.com/data/cirtms/lsat.dta
  obs:         1,000
 vars:             6                          3 Oct 2016 11:49
 size:        24,000
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              float   %9.0g
item1           float   %9.0g
item2           float   %9.0g
item3           float   %9.0g
item4           float   %9.0g
item5           float   %9.0g
-------------------------------------------------------------------------------
Sorted by:
```

In this output, Stata informs us that there are 1,000 observations on 6 variables, with both these numbers correctly reflecting the sample size and number of items in the original data file (`lsat.dat` or `lsat.dta`, in the ASCII or Stata format, respectively). The variable or item order, from left to right in that file, is as the one of their listing from top to bottom in the first column of the presented output. In particular, after a subject identifier, denoted `id` as usual, the $k = 5$ dichotomous items follow, which are stored in regular precision. (The remaining columns of this output pertain to data storage and technical details that are not of particular relevance here.)

At this stage, it is important as a matter of routine to make sure that the number of observations and variables in the read data file, which are indicated at the top of the above output, indeed equal the sample size and number of variables in the original dataset, respectively. As mentioned earlier, these two numbers should be known beforehand to the researcher. That sample size and variable number check is accomplished by inspecting the corresponding two numbers in the second and third lines from the top of the output resulting from the Stata command `describe` (or `d`). If the number of observations or variables in the read-in data file does not match the sample size or the number of variables in the initial data file, the reason for this discrepancy needs to be found and corrected before proceeding any further with the next steps outlined below. (That is, an accessed data file with a discrepancy for either of these two numbers should not be processed any further or analyzed until this discrepancy is resolved and the original data file correctly read in with the software.)

A look next at the descriptive statistics of the items can also be informative, in particular prior to commencing the IRT analyses discussed in the next section 6.2. This is achieved with the following command (which can also be shortened to `su`):

```
. summarize item1-item5
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       item1 |      1,000        .924    .2651307          0          1
       item2 |      1,000        .709    .4544508          0          1
       item3 |      1,000        .553    .4974318          0          1
       item4 |      1,000        .763    .4254551          0          1
       item5 |      1,000         .87    .3364717          0          1
```

By examining the mean of each item, which in this dichotomous item case equals the proportion of "correct" responses, we see from the last presented output that all items are associated with higher than 0.5 probability for that response (denoted "1"). The highest variance is exhibited by item 3, which is because its probability of "correct" response is closest to 0.5 (for example, Raykov and Marcoulides [2011]).

With this initial examination of the dataset of interest, we are now ready to proceed to fitting logistic IRT models and interpreting associated results in the next section. For the specific aims of the present empirical illustration using the LSAT dataset, we assume that each of the five items in it measures a particular aspect of the trait (construct) general mental ability.

6.2 Fitting a two-parameter logistic model

As discussed in chapter 5, the 2PL model represents a general logistic model for empirical settings with no guessing. (We assume initially no guessing on any item in a given set or instrument of interest and will revisit this matter in a following section as well as later in the book.) Therefore, we commence our IRT modeling effort by fitting this model. To this end, we use the following Stata command:

```
. irt 2pl item1-item5
```

This straightforward request yields this output:

```
Fitting fixed-effects model:
Iteration 0:   log likelihood = -2504.5114
Iteration 1:   log likelihood = -2493.5307
Iteration 2:   log likelihood = -2493.4367
Iteration 3:   log likelihood = -2493.4367
Fitting full model:
Iteration 0:   log likelihood = -2478.6867
Iteration 1:   log likelihood = -2467.6539
Iteration 2:   log likelihood = -2466.6565
Iteration 3:   log likelihood = -2466.6536
Iteration 4:   log likelihood = -2466.6536
Two-parameter logistic model                    Number of obs     =      1,000
Log likelihood = -2466.6536
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
item1						
Discrim	.8256703	.2581376	3.20	0.001	.3197299	1.331611
Diff	-3.358777	.8665242	-3.88	0.000	-5.057133	-1.660421
item2						
Discrim	.7227513	.1866698	3.87	0.000	.3568852	1.088618
Diff	-1.370049	.307467	-4.46	0.000	-1.972673	-.7674249
item3						
Discrim	.8907338	.2326049	3.83	0.000	.4348366	1.346631
Diff	-.2796988	.0996259	-2.81	0.005	-.4749621	-.0844356
item4						
Discrim	.6883831	.1851495	3.72	0.000	.3254968	1.05127
Diff	-1.866349	.4343093	-4.30	0.000	-2.71758	-1.015118
item5						
Discrim	.6568946	.2099182	3.13	0.002	.2454624	1.068327
Diff	-3.125751	.8711505	-3.59	0.000	-4.833174	-1.418327

As indicated at the top of this output, after several iterations, the underlying numerical optimization procedure converged to the solution provided above. A main index of relative model fit, which will be used later for comparing the 2PL model with the 1PL model and a 3PL model fit to the same dataset, is the maximized log-likelihood. Its value may be treated, somewhat informally, as a "goodness-of-fit" measure or index that does not explicitly account for model complexity. Hence, this index is best used for model comparison. (For a more detailed discussion, see section 6.5 dealing with nested models and chapter 7 for the data likelihood concept and its maximization.) We note that in the present example, the maximized log-likelihood equals -2466.65 (rounded off).

For each of the five analyzed items, the discrimination and difficulty parameter estimates follow in the subsequently presented panel of the above output. In addition, their associated standard errors, test statistics for being equal to 0 in the population of concern, and pertinent p-values, as well as 95% confidence intervals (CIs), are listed in the remainder of the corresponding rows. Thereby, the information pertaining to the a parameter (item discrimination) precedes that for the b parameter (item difficulty) for each item in this default output layout. (This result presentation layout may arguably be more often of interest in empirical research, but alternative ones are also available; see below.) These findings suggest that under the 2PL model, each item has nonzero discrimination and difficulty parameters in the studied subject population. This interpretation is based on direct inspection of the last 3 columns of above output and, in particular, their CIs, which do not contain the 0 point (suggesting none of these 10 parameters are 0 in the studied population, which as indicated earlier is of relevance with respect to the item discrimination parameters but not the item difficulty parameters).

If one wished a different solution presentation, reordering the lines of the last output is also possible. For instance, if one desired to have the items first "ranked" in terms of their a parameters in ascending order, we request it in the following way:

```
. estat report, byparm sort(a)
Two-parameter logistic model                    Number of obs     =      1,000
Log likelihood = -2466.6536
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
item5	.6568946	.2099182	3.13	0.002	.2454624	1.068327
item4	.6883831	.1851495	3.72	0.000	.3254968	1.05127
item2	.7227513	.1866698	3.87	0.000	.3568852	1.088618
item1	.8256703	.2581376	3.20	0.001	.3197299	1.331611
item3	.8907338	.2326049	3.83	0.000	.4348366	1.346631
Diff						
item5	-3.125751	.8711505	-3.59	0.000	-4.833174	-1.418327
item4	-1.866349	.4343093	-4.30	0.000	-2.71758	-1.015118
item2	-1.370049	.307467	-4.46	0.000	-1.972673	-.7674249
item1	-3.358777	.8665242	-3.88	0.000	-5.057133	-1.660421
item3	-.2796988	.0996259	-2.81	0.005	-.4749621	-.0844356

We note from this output that the only effect of the last-used Stata command is the reordering of the rows in the earlier presented item results section. Hence, none of the results associated with the fit model are changed (because no new model has been fit to the same dataset analyzed). As we can see from the top panel of the last output, in the used sample (dataset), item 5 is the least discriminating one, as judged by the item discrimination parameter estimates. However, given the relatively large standard errors compared with the differences in these estimates for the other items, one cannot suggest from this observation only that item 5 would be the least discriminating item also in the population. In fact, keeping in mind the relatively sizable standard errors, one may as well suggest that the five items have very similar discrimination parameters. This is a potentially rather interesting relationship with respect to the studied population (see also chapter 5). We thus keep in mind this discrimination parameter similarity across items and will pursue it in more detail in the next section.

Alternatively, if one wished instead to have the item difficulty parameter estimates ranked in ascending order, we achieve it this way, observing also that they are all negative here (see lower panel of output below):

```
. estat report, byparm sort(b)
Two-parameter logistic model                    Number of obs     =      1,000
Log likelihood = -2466.6536
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim						
item1	.8256703	.2581376	3.20	0.001	.3197299	1.331611
item5	.6568946	.2099182	3.13	0.002	.2454624	1.068327
item4	.6883831	.1851495	3.72	0.000	.3254968	1.05127
item2	.7227513	.1866698	3.87	0.000	.3568852	1.088618
item3	.8907338	.2326049	3.83	0.000	.4348366	1.346631
Diff						
item1	-3.358777	.8665242	-3.88	0.000	-5.057133	-1.660421
item5	-3.125751	.8711505	-3.59	0.000	-4.833174	-1.418327
item4	-1.866349	.4343093	-4.30	0.000	-2.71758	-1.015118
item2	-1.370049	.307467	-4.46	0.000	-1.972673	-.7674249
item3	-.2796988	.0996259	-2.81	0.005	-.4749621	-.0844356

However, these 10 parameter estimates (of 5 item discrimination and 5 item difficulty parameters) are not often easy to interpret in purely numeric terms. Thus, we could for instance graph the corresponding item characteristic curves (ICCs) to aid with their interpretation. We achieve it with the following Stata command:

```
. irtgraph icc item1-item5
```

This yields the graph presented in figure 6.3 [with different colors assigned as software default to the different ICCs].

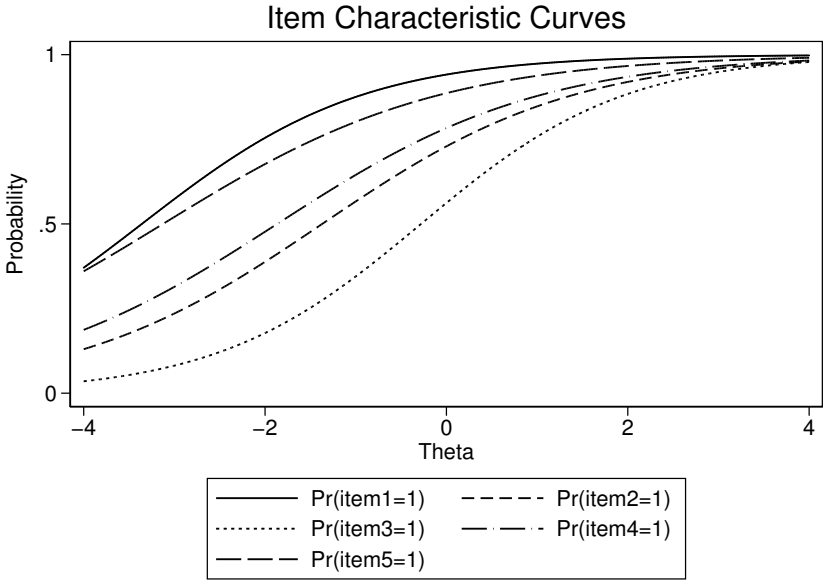

Figure 6.3. Graph of the item characteristic curves for the $k = 5$ analyzed items

We notice from figure 6.3 (see also last presented output) that item 1 is the easiest in the analyzed dataset. To observe this, imagine drawing a horizontal line at the point symbolizing the probability of 0.5 on the vertical axis. Because this line first crosses the ICC of item 1, as one moves from left to right on the horizontal axis, this item is easiest here (see also figure 6.4 below). In the same way, we can also observe that item 3 seems to be the hardest (in this dataset). This is because for item 3, the point on the horizontal axis that corresponds to the intersection of its ICC with that imaginary horizontal line at 0.5 probability is to the right of any such point for the remaining 4 items (see also figure 6.4). In addition, figure 6.3 suggests that the tangents to each ICC at its inflection point (the point of intersection of the ICC with that imaginary horizontal line at 0.5 probability) are possibly fairly close to parallel. Again, to be more confident in such an interpretation, we need additional analyses that we will conduct in section 6.3.

To obtain more precise graphical information about possible item difficulty differences, we can request pointing out the location of the difficulty parameter estimates on the ICC plot. We achieve this with the following command (note that it is the first ICC graphing command with an added subcommand stated after the comma, and see figure 6.4 for the resulting graph):