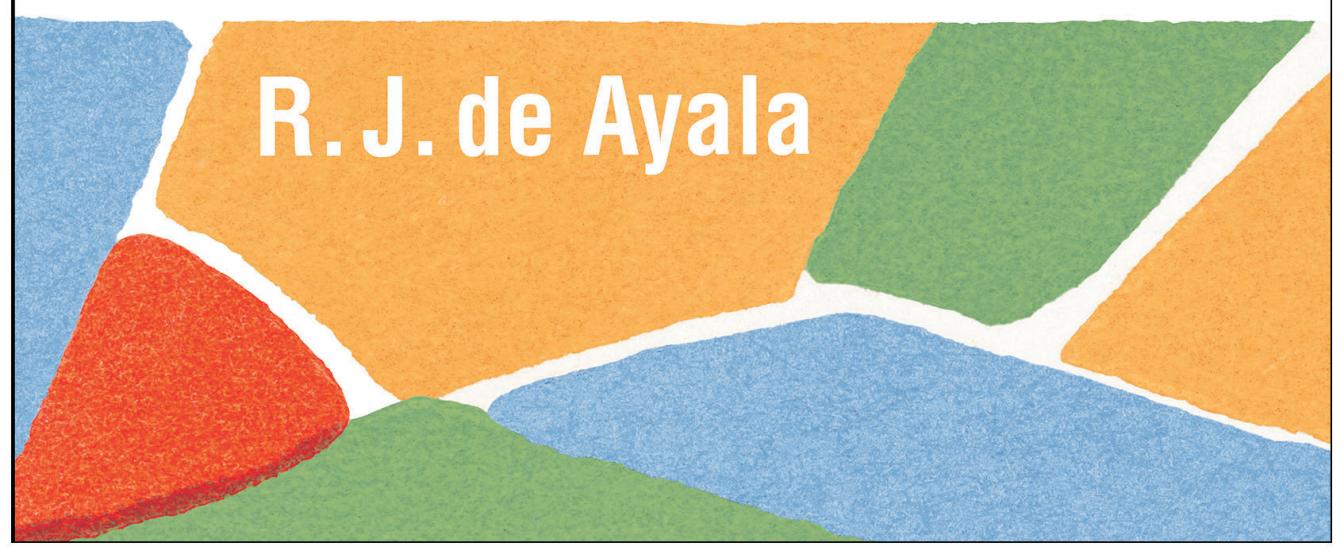


SECOND EDITION

THE THEORY
AND PRACTICE OF
ITEM RESPONSE
THEORY



R. J. de Ayala

The Theory and Practice of Item Response Theory

Methodology in the Social Sciences

David A. Kenny, Founding Editor

Todd D. Little, Series Editor

www.guilford.com/MSS

This series provides applied researchers and students with analysis and research design books that emphasize the use of methods to answer research questions. Rather than emphasizing statistical theory, each volume in the series illustrates when a technique should (and should not) be used and how the output from available software programs should (and should not) be interpreted. Common pitfalls as well as areas of further development are clearly articulated.

RECENT VOLUMES

PSYCHOMETRIC METHODS: THEORY INTO PRACTICE

Larry R. Price

MEASUREMENT THEORY AND APPLICATIONS FOR THE SOCIAL SCIENCES

Deborah L. Bandalos

CONDUCTING PERSONAL NETWORK RESEARCH: A PRACTICAL GUIDE

Christopher McCarty, Miranda J. Lubbers, Raffaele Vacca, and José Luis Molina

QUASI-EXPERIMENTATION: A GUIDE TO DESIGN AND ANALYSIS

Charles S. Reichardt

THEORY CONSTRUCTION AND MODEL-BUILDING SKILLS: A PRACTICAL GUIDE
FOR SOCIAL SCIENTISTS, SECOND EDITION

James Jaccard and Jacob Jacoby

LONGITUDINAL STRUCTURAL EQUATION MODELING WITH MPLUS:
A LATENT STATE-TRAIT PERSPECTIVE

Christian Geiser

COMPOSITE-BASED STRUCTURAL EQUATION MODELING: ANALYZING LATENT
AND EMERGENT VARIABLES

Jörg Henseler

BAYESIAN STRUCTURAL EQUATION MODELING

Sarah Depaoli

INTRODUCTION TO MEDIATION, MODERATION, AND CONDITIONAL
PROCESS ANALYSIS: A REGRESSION-BASED APPROACH, THIRD EDITION

Andrew F. Hayes

THE THEORY AND PRACTICE OF ITEM RESPONSE THEORY, SECOND EDITION

R. J. de Ayala

The Theory and Practice of Item Response Theory

SECOND EDITION

R. J. de Ayala

Series Editor's Note by Todd D. Little



THE GUILFORD PRESS
New York London

Copyright © 2022 The Guilford Press
A Division of Guilford Publications, Inc.
370 Seventh Avenue, Suite 1200, New York, NY 10001
www.guilford.com

All rights reserved

No part of this book may be reproduced, translated, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission from the publisher.

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication

Names: De Ayala, R. J. (Rafael Jaime), 1957– author. | Little, Todd D., editor.

Title: The theory and practice of item response theory / R. J. de Ayala ; series editor's note by Todd D. Little.

Description: Second edition. | New York : The Guilford Press, [2022] |

Series: Methodology in the social sciences | Includes bibliographical references and indexes.

Identifiers: LCCN 2021044937 | ISBN 9781462547753 (cloth)

Subjects: LCSH: Item response theory. | Social sciences—Mathematical models. | Social sciences—Statistical methods. | Psychometrics. |

BISAC: SOCIAL SCIENCE / Statistics | BUSINESS & ECONOMICS / Statistics

Classification: LCC H61.25 .D4 2022 | DDC 300.1/5118—dc23/eng/20211206

LC record available at <https://lccn.loc.gov/2021044937>

A mi esposa, Stephanie, y mi hija, Isabel

Series Editor's Note

Although latent variable modeling provides an amazingly powerful set of modeling tools, it isn't a panacea for bad measurement. Measurement is the bedrock of any statistical modeling endeavor. As E. L. Thorndike once said back in 1918, "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." Here, item response theory (IRT) is one of the essential tools to probe the inherent meaning of responses when items are categorical in nature and glean the underlying latent trait score of the respondents.

de Ayala's second edition brings so much added value to what was already a go-to resource for learning the foundations of IRT. His approach is not a cookbook of "what" to do, but instead his focus is on "how" to utilize the various tools from the world of IRT with the right amount of detail to know "why." In the body of each chapter, you'll find the practical guidance needed and how to implement and interpret IRT-based models. The new material provides accessible technical details for understanding how the engine under the hood of an IRT model actually works. The family of IRT models and their variety of options are brought to you across a well-chosen set of software platforms, from free (e.g., R-based packages such as `lme4`, `mirt`, and `mixRasch`) to proprietary (e.g., SAS, SPSS, and flexMIRT). In fact, each chapter contains both R and non-R software examples!

The utility of IRT modeling goes well beyond its roots in educational measurement. de Ayala does a wonderful job of using general terminology that broadens the practical reach and impact that IRT modeling can have for researchers across the spectrum of the social and behavioral sciences. Core ideas such as measurement equating and measurement equivalence are discussed with an eye to the myriad research applications where IRT-based models are needed (e.g., item banking, test equating, certification exams, adaptive testing, protocol development). His model data-fit approach ties the statistical machinery with practical need, which yields a refreshingly comprehensive yet accessible coverage of all IRT models, from established (e.g., Rausch) to cutting edge (e.g., multi-level IRT, which is new to this edition). Examples don't bounce around, so you can focus on what each model can reveal and understand how to interpret the results.

With de Ayala's rich resource as your guide, you'll be at the forefront of best practices in IRT modeling.

With the measurement precision that IRT provides, your research will become unequivocally better.

With better research come answers that truly impact the policies and practices in arenas spanning public health, behavioral sciences, social sciences, and, of course, education.

As always, enjoy! You'll be gratified with de Ayala's guide to all things IRT.

TODD D. LITTLE

*Taking a break to go fishing
Melrose, Montana*

Preface

This edition of *The Theory and Practice of Item Response Theory* includes new material on multilevel and loglinear models, updates the fit statistics that are discussed, and includes the use of R, SAS, and flexMIRT while removing some software used in the first edition. Item response theory (IRT) methodology forms the foundation of many psychometric applications. For instance, IRT is used for equating alternate examination forms, for instrument design and development, for item banking, for computerized adaptive testing, for certification testing, and in various other psychological assessment domains. Well-trained psychometricians are increasingly expected to understand and be able to use IRT.

In this book we address the “how to” of applying IRT models while at the same time providing enough technical substance to answer the “why” questions. We make extensive use of endnotes and appendices to address some technical issues, and the reader is referred to appropriate sources for even greater technical depth than is provided here. To facilitate understanding the application of the models, we use common data sets across the chapters. In addition, the exemplary model applications employ several common software packages. Some of these packages are free (BIGSTEPS, NOHARM, Facets, R packages [`lme4`, `mirt`, `mixRasch`, `PerFit`, `psychomix`, `sirt`]), while others are commercially available (BILOG-MG, flexMIRT, SAS [`proc IRT` and `proc glimmix`], WINMIRA, SPSS, and SYSTAT). For each chapter, examples use R and non-R software. The terminology used in this book is more general than is typically seen in IRT textbooks and is not specific to educational measurement. The reader is assumed to be familiar with common psychometric concepts such as reliability, validity, scaling, levels of measurement, and factor analysis, as well as regression analysis.

Chapter 1 “begins at the beginning” by presenting the basic concept of why we are interested in performing measurement. There is a short philosophical treatment of measurement, pointing out the desirable characteristics that we would like measurement to have. The traditional psychometric concepts of levels of measurement, reliability, validity, and various approaches to measurement, such as IRT, latent class analysis, and classical test theory, are woven into this introduction.

Chapter 2 introduces the simplest IRT model, the Rasch or one-parameter logistic (1PL) model. This model establishes the principles that underlie the more complex models discussed in subsequent chapters. The chapter also introduces the empirical data set that is used in the next four chapters and discusses philosophical differences between Rasch and non-Rasch models. Chapters 3 and 4 conceptually present two different parameter estimation techniques that are applied to the data set introduced in Chapter 2. Chapter 3's application begins by examining the unidimensionality assumption through nonlinear factor analysis and proceeds to item parameter estimation and interpretation. In Chapter 4 we reanalyze our data using a different estimation technique and a different program. In addition, we present alternative item- and model-level fit analyses by using statistical and graphical methods.

Chapter 5 presents the two-parameter logistic (2PL) model. Building on the 1PL model presentation, the chapter is concerned only with the features that are unique to this model. Additional fit analysis methods are introduced in this chapter in our reanalysis of the data used with the 1PL model. Similarly, Chapter 6 introduces the characteristics specific to the three-parameter logistic model and reanalyzes the data used in the 1PL and 2PL model examples. By the end of Chapter 6, our fit analysis has evolved to include both item- and model-level fit analyses, model comparison statistics, and person fit analysis. The chapter also examines IRT's assumptions regarding functional form and conditional independence. As such, through these chapters we demonstrate the steps of model and data fit that would normally be used in practice.

Because data are not always dichotomous, we next discuss models that are appropriate for polytomous data (e.g., data from Likert response scales) in Chapters 7–9. The models are divided between those for ordered and those for unordered polytomous data. Chapters 7 and 8 address ordered polytomous data from both Rasch and non-Rasch perspectives (i.e., Rasch: partial credit and rating scale models; non-Rasch: generalized partial credit and graded response models). As in other parts of the book, common data sets are used across these chapters. Modeling unordered polytomous data as well as simultaneously using multiple IRT models are addressed in Chapter 9.

All of the models presented to this point assume that an individual's responses are a function of a single-person latent variable. In Chapter 10 we generalize the models presented in Chapters 5 and 6 to multiple latent variables.

As mentioned, IRT is used for equating multiple forms and in the creation of item banks. Techniques for accomplishing both of these purposes are presented in Chapter 11. Our final addition to our model–data fit toolbox is provided in Chapter 12, where we are concerned with whether an item functions differently for different groups of people. Stated another way, do factors (e.g., the respondents' gender) that are tangential to the construct an item is supposed to be measuring affect how people respond to the item? If so, then the item may be biased against those individuals. Techniques for identifying items with such discrepancies are presented in this chapter. Finally, Chapter 13 covers multilevel IRT models. The examples demonstrate two- and three-level models.

Appendices A–G are not in the book but can be accessed at the companion website (www.guilford.com/deayala-materials). There you will also find a link to data, syntax, and output files in different software packages for the book's examples.

I would like to acknowledge that this book has been influenced by my interactions with various individuals over the past two decades. Barbara Dodd, Bill Koch, Earl Jennings, Chan Dayton, Bill Hays, Frank Baker, Mark Reckase, Dave Weiss, and Seock-Ho Kim are some of them. I am grateful for the thoughtful suggestions of the reviewers, whose identity and comments were blinded until this second edition was completed: Ojmarrh Mitchell, School of Criminology and Criminal Justice, Arizona State University; Karen M. Schmidt, Department of Psychology, University of Virginia; Michael D. Toland, Educational Counseling, University of Kentucky; Leigh Wang, Education, University of Cincinnati; and, Larry R. Price, Director of Methodology, Measurement, and Statistical Analysis, Texas State University. My apologies to anyone whom I have omitted. I would also like to acknowledge the graduate students in psychometrics whom I have had the pleasure of teaching. I am appreciative of the support and patience of C. Deborah Laughton, Research Methods and Statistics Publisher at The Guilford Press.

Contents

Symbols and Acronyms

xxi

1 • Introduction to Measurement

1

- Measurement / 1
- Some Measurement Issues / 3
- Item Response Theory / 5
- Classical Test Theory / 5
- Latent Class Analysis / 8
- Summary / 9

2 • The One-Parameter Model

12

- Conceptual Development of the Rasch Model / 12
- The One-Parameter Model / 17
- The One-Parameter Logistic Model and the Rasch Model / 20
- Assumptions Underlying the Model / 21
- An Empirical Data Set: The Mathematics Data Set / 23
- Conceptually Estimating an Individual's Location / 23
- Some Pragmatic Characteristics of Maximum Likelihood Estimates / 28
- The Standard Error of Estimate and Information / 29
- An Instrument's Estimation Capacity / 32
- Summary / 35

3 • Joint Maximum Likelihood Parameter Estimation

42

- Joint Maximum Likelihood Estimation / 42
- Indeterminacy of Parameter Estimates / 44
- How Large a Calibration Sample? / 45
- Example: Application of the Rasch Model to the Mathematics Data, JMLE, BIGSTEPS / 46*
- Example: Application of the Rasch Model to the Mathematics Data, JMLE, mixRasch / 68*
- Validity Evidence / 75

Summary of the Application of the Rasch Model / 76	
Summary / 77	
4 • Marginal Maximum Likelihood Parameter Estimation	86
Marginal Maximum Likelihood Estimation / 86	
Estimating an Individual's Location: Expected A Posteriori / 93	
<i>Example: Application of the Rasch Model to the Mathematics Data, MMLE, BILOG-MG / 98</i>	
Metric Transformation and the Total Characteristic Function / 111	
<i>Example: Application of the Rasch Model to the Mathematics Data, MMLE, mirt / 115</i>	
Summary / 125	
5 • The Two-Parameter Model	135
Conceptual Development of the Two-Parameter Model / 135	
Information for the Two-Parameter Model / 137	
Conceptual Parameter Estimation for the 2PL Model / 139	
How Large a Calibration Sample? / 140	
Metric Transformation, 2PL Model / 142	
<i>Example: Application of the 2PL Model to the Mathematics Data, MMLE, BILOG-MG / 143</i>	
Fit Assessment: An Alternative Approach for Assessing Invariance / 146	
<i>Example: Application of the 2PL Model to the Mathematics Data, MMLE, mirt / 152</i>	
Information and Relative Efficiency / 162	
Summary / 165	
6 • The Three-Parameter Model	179
Conceptual Development of the Three-Parameter Model / 179	
Additional Comments about the Pseudo-Guessing Parameter, χ_j / 182	
Conceptual Parameter Estimation for the 3PL Model / 183	
How Large a Calibration Sample? / 187	
Assessing Conditional Independence / 188	
<i>Example: Application of the 3PL Model to the Mathematics Data, MMLE, BILOG-MG / 192</i>	
Fit Assessment: Conditional Independence Assessment / 195	
Fit Assessment: Model Comparison / 198	
<i>Example: Application of the 3PL Model to the Mathematics Data, MMLE, mirt / 200</i>	
Assessing Person Fit: Appropriateness Measurement / 209	
Information for the Three-Parameter Model / 216	
Metric Transformation, 3PL Model / 220	
Handling Missing Responses / 220	
Issues to Consider in Selecting among the 1PL, 2PL, and 3PL Models / 224	
Summary / 226	

7 • Rasch Models for Ordered Polytomous Data	237
Conceptual Development of the Partial Credit Model / 238	
Conceptual Parameter Estimation of the PC Model / 243	
<i>Example: Application of the PC Model to a Reasoning Ability Instrument, MMLE, flexMIRT / 244</i>	
<i>Example: Application of the PC Model to a Reasoning Ability Instrument, MMLE, mirt / 256</i>	
The Rating Scale Model / 267	
Conceptual Parameter Estimation of the RS Model / 272	
<i>Example: Application of the RS Model to an Attitudes Toward Condoms Scale, JMLE, BIGSTEPS / 272</i>	
<i>Example: Application of the PC Model to an Attitudes Toward Condoms Scale, JMLE, mixRasch / 287</i>	
How Large a Calibration Sample? / 292	
Information for the PC and RS Models / 294	
Metric Transformation, PC and RS Models / 296	
Summary / 296	
8 • Non-Rasch Models for Ordered Polytomous Data	313
The Generalized Partial Credit Model / 313	
<i>Example: Application of the GPC Model to a Reasoning Ability Instrument, MMLE, flexMIRT / 318</i>	
<i>Example: Application of the GPC Model to a Reasoning Ability Instrument, MMLE, mirt / 321</i>	
Conceptual Development of the Graded Response Model / 324	
How Large a Calibration Sample? / 333	
Information for Graded Data / 334	
Metric Transformation, GPC and GR Models / 336	
<i>Example: Application of the GR Model to an Attitudes Toward Condoms Scale, MMLE, flexMIRT / 337</i>	
<i>Example: Application of the GR Model to an Attitudes Toward Condoms Scale, MMLE, mirt / 340</i>	
Conceptual Development of the Continuous Response Model / 343	
Summary / 351	
9 • Models for Nominal Polytomous Data	356
Conceptual Development of the Nominal Response Model / 357	
Information for the NR Model / 365	
Metric Transformation, NR Model / 366	
Conceptual Development of the Multiple-Choice Model / 366	
How Large a Calibration Sample? / 368	
<i>Example: Application of the NR Model to a General Science Test, MMLE, mirt / 370</i>	
Summary / 383	

10 • Models for Multidimensional Data	391
Conceptual Development of a Multidimensional IRT Model / 391	
Multidimensional Item Location and Discrimination / 397	
Item Vectors and Vector Graphs / 401	
The Multidimensional Three-Parameter Logistic Model / 404	
Assumptions of the MIRT Model / 404	
Estimation of the M2PL Model / 405	
Information for the M2PL Model / 406	
Indeterminacy in MIRT / 408	
Metric Transformation, M2PL Model / 410	
<i>Example: Calibration of Interpersonal Engagement Instrument, M2PL Model, sirt.noharm / 411</i>	
Obtaining Person Location Estimates / 421	
<i>Example: Calibration of Interpersonal Engagement Instrument, M2PL Model, mirt / 422</i>	
<i>Example: Calibration of Interpersonal Engagement Instrument, M2PL Model, flexMIRT / 429</i>	
Summary / 431	
11 • Linking and Equating	443
Equating Defined / 443	
Equating: Data Collection Phase / 445	
Equating: Transformation Phase / 446	
<i>Example: Application of the Total Characteristic Function Equating Method, EQUATE / 454</i>	
<i>Example: Application of the Total Characteristic Function Equating Method, SNSequate / 463</i>	
<i>Example: Fixed-Item and Concurrent Calibration Equating / 465</i>	
Summary / 471	
12 • Differential Item Functioning	478
Differential Item Functioning and Item Bias / 479	
Mantel–Haenszel Chi-Square / 483	
The TSW Likelihood Ratio Test / 486	
Logistic Regression / 487	
<i>Example: DIF Analysis of Vocabulary Test, SAS CMH / 491</i>	
<i>Example: DIF Analysis of Vocabulary Test, mantelhaen.test and difR / 494</i>	
<i>Example: DIF Analysis of Vocabulary Test, SAS proc logistic / 501</i>	
<i>Example: DIF Analysis of Vocabulary Test, glm and difR / 508</i>	
Summary / 518	

13 • Multilevel IRT Models 525

Multilevel IRT—Two Levels / 525

*Example: Estimating the Rasch Model from a Multilevel Perspective,
proc glimmix / 530*

Example: Rasch Model Estimation, lme4 / 541

Person-Level Predictors for Items / 545

*Example: Person-Level Predictors for Items—DIF Analysis,
proc glimmix / 547*

Example: Person-Level Predictors for Items—DIF Analysis, lme4 / 551

Person-Level Predictors for Respondents / 556

*Example: Person-Level Predictors for Respondents—Nutrition Literacy,
proc glimmix / 558*

Example: Person-Level Predictors for Respondents, lme4 / 562

Item-Level Predictors for Items / 567

*Example: Item-Level Predictors for Items—Nutrition Literacy,
proc glimmix / 569*

Example: Item-Level Predictors for Items—Nutrition Literacy, lme4 / 571

Multilevel IRT—Three Levels / 574

*Example: Three-Level Model Analysis—Nutrition Literacy,
proc glimmix / 579*

Example: Three-Level Analysis of Nutrition Literacy Data, lme4 / 582

Summary / 587

Appendices A–G can be accessed online at the book's companion website (www.guilford.com/deayala-materials), which also provides links to data, syntax, and output files in different software packages for the book's examples.

Appendix A. Maximum Likelihood Estimation of Person Locations

Estimating an Individual's Location: Empirical Maximum Likelihood
Estimation

Estimating an Individual's Location: Newton's Method for MLE

R Function for MLE of θ with the Rasch Model

Revisiting Zero Variance Binary Response Patterns

Appendix B. Maximum Likelihood Estimation of Item Locations

R function for MLE of δ with the Rasch Model

Appendix C. The Normal Ogive Models

Conceptual Development of the Normal Ogive Model

The Relationship between IRT Statistics and Traditional Item Analysis
Indices

Relationship of the Two-Parameter Normal Ogive and Logistic Models

Extending the Two-Parameter Normal Ogive Model to a Multidimensional
Space

Appendix D. Computerized Adaptive Testing

A Brief History

Fixed-Branching Techniques

Variable-Branching Techniques

Advantages of Variable-Branching over Fixed-Branching Methods
IRT-Based Variable-Branching Adaptive Testing Algorithm
Appendix E. Linear Logistic Test Model (LLTM)
Example of LLTM Calibration Using eRm
Appendix F. Mixture Models
Latent Class Analysis
Mixture Rasch Model
Example: Application of the Mixture Rasch Model to Writing Problem Data, CMLE, WINMIRA
Example: Application of the Mixture Rasch Model to Writing Problem Data, CMLE, psychomix
Appendix G. Miscellanea
Using Principal Axis for Estimating Item Discrimination
Infinite Item Discrimination Parameter Estimates
Example: NOHARM Unidimensional Calibration
An Approximate Chi-Square Statistic for NOHARM
Relative Efficiency, Monotonicity, and Information
FORTRAN Formats
Odds, Odds Ratios, and Logits
The Person Response Function
Linking: A Temperature Analogy Example
Should DIF Analyses Be Based on Latent Classes?
The Separation and Reliability Indices
Dependency in Traditional Item Statistics and Observed Scores
Conditional Independence Using Q_3
Standalone NOHARM Calibration of Interpersonal Engagement Instrument, M2PL Model
CFI, GFI, M_2 , RMSEA, TLI, and SRMR
An Introduction to Kernel Equating
Correspondence between the Rasch Model and a Loglinear Model
R Introduction

References	597
Author Index	625
Subject Index	631
About the Author	643

Symbols and Acronyms

INDICES

F	number of factors	$f = 1, \dots, F$
L	number of items	$j = 1, \dots, L$
N	number of persons	$i = 1, \dots, N$
m	number of response categories	$k = 1, \dots, m$
m	number of operations, thresholds, boundaries	$k = 1, \dots, m$ or $k = 0, \dots, m$
R	number of quadrature points	$r = 1, \dots, R$
S	number of cognitive operations	$s = 1, \dots, S$
\underline{x}	response vector	
x_j	dichotomous response or category score on item j	
x_{jk}	polytomous response on item j 's k th category	
ν	index for latent classes (nu)	$\nu = 1, \dots, G$
π_ν	latent class ν proportion	

SYMBOLS

\underline{w}	eigenvector	λ	eigenvalue—no superscript; with superscripts: a loglinear effect (lambda)
Σ	variance/covariance matrix	p	probability
h^2	communality	a	item loading
$\underline{\upsilon}$	population parameters vector; MMLE (upsilon)	$\underline{\Omega}$	item parameter matrix
X_I	observed score person i	P_j	item traditional difficulty, proportion item j correction
q_j	item j score		
Γ	DIF, group variable (capital gamma)	Λ	DIF, person location (e.g., θ or X_i) (capital lambda)
L	likelihood function	T	true or “trait” score (capital tau)
$\ln L$	log likelihood function	\mathcal{ET}	expected proportion (of 1s) trait score
\mathcal{E}	expectation	E_i	error score (capital epsilon)
$I(\theta)$	total information	$I_j(\theta)$	item information

$I_{\omega}(\Theta)$	multidimensional total information	$I_{j\omega}(\Theta)$	multidimensional item information
\underline{W}	information matrix, nominal response model	Φ	cumulative normal distribution (capital phi)
\prod	product operator or symbol	D	scaling constant
Ξ	convergence criterion (capital xi)		

PARAMETERS

θ	person location (theta)	τ	threshold (tau)
α or α_j	item discrimination (alpha)		
δ_j	item j location (delta)	δ_{jh} or δ_{jk}	item j transition location
		δ_{x_j}	category boundary location
χ_j	pseudo-guessing (lower asymptote) (lowercase chi)	Υ_j	item j upper asymptote (uppercase epsilon)
γ_j	intercept (gamma)	$\delta_{k jv}$	conditional probability (LCA)
ϕ	“don’t know” category (phi)		
Δ_j	multidimensional item location (capital delta)	A_{ω}	multidimensional directional discrimination (capital alpha)
A_j	multidimensional item discrimination (capital alpha)	ω	angle or direction in the multidimensional space (omega)
η_s	elementary component s		
$\sigma_e(\hat{\theta})$	standard error of person location	$s_e(\hat{\theta})$	sample standard error
$\sigma_e(\hat{\delta})$	standard error item location	$s_e(\hat{\delta})$	sample standard error
σ_{meas}	standard error of measurement	ϑ_{ij}	logit person i and item j
$\beta_{(0)i}$	constant (level-1) for person i	$\beta_{(q)i}$	q th regression (slope) coefficient for person i and item j
$X_{(q)ij}$	q th indicator predictor (dummy variable) for person i and item j	$\theta_{(0)i}^r$	random person effect for person
ζ_{00}	constant (level-2)	$\zeta_{(q)1}$	q th regression (level-2) coefficient for first predictor (level-2)

The estimate of a parameter is represented with a circumflex. For example, the estimate of the parameter θ is symbolized as $\hat{\theta}$, the estimate of the parameter α is represented as $\hat{\alpha}$, etc.

TRANSFORMATION EQUATIONS

$$\begin{aligned}\xi^* &= \zeta(\xi) + \kappa & \gamma_j^* &= \gamma_j - \frac{\alpha_j(\kappa)}{\zeta} & \alpha_j^* &= \frac{\alpha_j}{\zeta} \\ \underline{\theta}_i^* &= \underline{Z}\underline{\theta}_i + \underline{\kappa} & \gamma_j^* &= \gamma_j - \underline{\alpha}'_j \underline{Z}^{-1} \underline{\kappa} & \underline{\alpha}_j^* &= (\underline{Z}^{-1})' \underline{\alpha}_j\end{aligned}$$

κ and ζ , metric transformation coefficients or equating coefficients; ξ is either δ_j or θ_i (κ : kappa; ζ : zeta; ξ : lowercase xi, pronounced “kas-eye”).

ACRONYMS

CTT	classical test theory
<i>df</i>	degrees of freedom
DIF	differential item functioning
EAP	expected a posteriori
GUI	graphical user interface
IRF	item response function
IRS	item response surface
IRT	item response theory (latent trait theory)
JMLE	joint maximum likelihood estimate or unconditional maximum likelihood estimation
LCA	latent class analysis
lnL	log likelihood
LSA	latent structure analysis
MAP	maximum a posteriori
MINF	multidimensional information
MIRT	multidimensional item response theory
MLE	maximum likelihood estimation
MMLE	marginal maximum likelihood estimation
<i>SD</i>	standard deviation
SEE	standard error of estimate
TCC	total characteristic curve or test characteristic curve
TCF	total characteristic function or test characteristic function

IRT MODELS (NONEXHAUSTIVE)

1PL	one-parameter logistic model
2PL	two-parameter logistic model

3PL	three-parameter logistic model
CR	continuous response model
GPC	generalized partial credit model (2PPC)
GR	graded response model
LLTM	linear logistic test model
M2PL	multidimensional compensatory two-parameter logistic model
M3PL	multidimensional compensatory three-parameter logistic model
MFRM	Many-Facet Rasch Model
MC	multiple-choice model
	mixture Rasch model
NR	nominal response model
PC	partial credit model

The Theory and Practice of Item Response Theory

1

Introduction to Measurement

I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the state of science, whatever the matter may be.

—Sir William Thomson (Lord Kelvin) (1891, p. 80)

This book is about a particular measurement perspective called *item response theory* (IRT), *latent trait theory*, or *item characteristic curve theory*. To understand this measurement perspective, we need to address what we mean by the concept of measurement. Measurement can be defined in many different ways. A classic definition is that measurement is “the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement” (Stevens, 1946, p. 677). Although commonly used in introductory measurement and statistics texts, this definition reflects a rather limited view. Measurement is more than just the assignment of numbers according to rules (i.e., labeling); it is a process by which an attempt is made to understand the nature of a variable (cf. Bridgman, 1928). Moreover, whether the process results in numeric values with inherent properties or the identification of different classes depends on whether we conceptualize the variable of interest as continuous or categorical. IRT provides one particular mathematical technique for performing measurement in which the variable is considered to be continuous in nature.

Measurement

For a simple example of measurement as a process, imagine that a researcher is interested in measuring generalized anxiety. Anxiety may be loosely defined as feelings that may range from general uneasiness to incapacitating attacks of terror. Because the very

nature of anxiety involves feelings, it is not possible to directly observe anxiety. As such, anxiety is an unobservable or *latent* variable or construct.

The measurement process involves deciding whether our latent variable, anxiety, should be conceptualized as categorical, continuous, or both. In the categorical case, we would classify individuals into qualitatively different latent groups so that, for example, one group may be interpreted as representing individuals with incapacitating anxiety and another group as representing individuals without anxiety. In this conceptualization, the persons differ from one another in kind on the latent variable. Typically, these latent categories are referred to as latent classes. Latent class analysis can be used to model this conceptualization of the data.

Alternatively, anxiety could be conceptualized as continuous. From this perspective, individuals differ from one another in their *quantity* of the latent variable. Thus, we might label the ends of the *latent continuum* as, say, “high anxiety” and “low anxiety.” Latent trait theory (a.k.a., item response theory) can be used to model this conceptualization of the data.

When the latent variable is conceptualized as having categorical *and* continuous facets, then we have a combination of two or more latent classes and one or more latent continua. In this case, the latent classes are subpopulations that are homogeneous with respect to the variable of interest but differ from one another in kind. Within each of these classes there is a latent continuum on which the individuals within the class may be located. Thus, we have a combination of latent classes and latent continua. For example, assume that our sample of respondents consists of two classes. One class could consist of individuals whose anxiety is so severe that they suffer from incapacitating attacks of terror. As such, these individuals are so qualitatively different from other persons that they need to be addressed separately from those whose anxiety is not so severe. Therefore, the second class contains individuals who do not suffer from incapacitating attacks of terror. Within each of these classes we have a latent continuum on which we locate the class’s respondents. Mixture models (e.g., the mixture Rasch model) can be used to model this conceptualization of the data.

Although we cannot observe our latent variable, its existence may be inferred from behavioral manifestations or *manifest* variables (e.g., restlessness, sleeping difficulties, headaches, trembling, muscle tension, item responses, self-reports). These manifestations allow for several different approaches to measuring generalized anxiety. For example, one approach may involve physiological assessment via an electromyogram of the degree of muscle tension. Other approaches might involve recording the number of hours spent sleeping or the frequency and duration of headaches, using a galvanic skin response (GSR) feedback device to assess sweat gland activity, or more psychological approaches, such as asking a series of questions. These approaches, either individually or collectively, provide our *operational definition* of generalized anxiety (Bridgman, 1928). That is, our operational definition specifies how we go about collecting our observations (i.e., the latent variable’s manifestations). Stated concisely, our interest is in our latent variable, and its operational definition is a means to that end.

The measurement process, so far, has involved our conceptualization of the latent variable’s nature and its operational definition. We also need to decide on the correspon-

dence between our observations of the individuals' anxiety levels and their locations on the continuum and/or in a class. In general, *scaling* is the process of establishing the correspondence between the observation data and the persons' locations on the latent variable. Once we have our individuals located on the latent variable, we can then compare them to one another. IRT is one approach to establishing this correspondence between the observation data and the persons' locations on the latent variable. Examples of other relevant scaling processes are Guttman Scalogram analysis (Guttman, 1950), Coombs Unfolding (Coombs, 1950), and the various Thurstone approaches (Thurstone, 1925, 1928, 1938). Alternative scaling approaches may be found in Dunn-Rankin, Knezeck, Wallace, and Zhang (2004), Gulliksen (1987), Maranell (1974), and Nunnally and Bernstein (1994).

Some Measurement Issues

Before proceeding to discuss various latent variable methods for scaling our observations, we need to discuss four issues. The first issue involves the *consistency of the measures*. By way of analogy, assume that we are measuring the length of a box. If our repeated measurements of the length of the box were constant, then these measurements would be considered to be highly consistent or to have high *reliability*. However, if these repeated measurements varied wildly from one another, then they would be considered to have low consistency or to have low reliability. In the former case, our measurements would have a small amount of error, whereas in the latter they would have a comparatively larger amount of error. The consistency (or lack thereof) would affect our confidence in the measurements. That is, in the first scenario we would have greater confidence in our measurements than in the second scenario.

The second issue concerns the *validity* of the measures. Although there are various types of validity, we define validity as the degree to which our measures are actually manifestations of the latent variable. As a contrarian example, assume we use the "frequency and duration of headaches" approach for measuring anxiety. Although some persons may recognize that there might be a relationship between "frequency and duration of headaches" and anxiety level, they may not consider this approach, in and of itself, to be an accurate "representation" of anxiety. In short, simply because we construct a measure does not mean the measure necessarily results in an accurate reflection of the variable of theoretical interest (i.e., our measurements may or may not have validity). A necessary, but not sufficient condition for our measurements to have validity is that they possess a high degree of reliability. Therefore, it is necessary to be concerned not only with the consistency of our measurements, but also with their validity. Obtaining validity evidence is part of the measurement process.

The third issue concerns a desirable property we would like our measurements to possess. Thurstone (1928) noted that a measuring instrument must not be seriously affected in its measuring function by the object of measurement. In other words, we would like our measurement instrument to be independent of what it is we are measuring. If this is true, then the instrument possesses the property of *invariance*. For

instance, if we measure the size of a shoe box by using a meter stick, then the measurement instrument (i.e., the meter stick) is not affected by and is independent of which box is measured. Contrast this with the situation in which measuring a shoe box's size is done not by using a meter stick, but by stretching a string along the shortest dimension of the box and cutting the string so that its length equals the shortest dimension. This string would serve as our measurement instrument, and we would use it to measure the other two dimensions of the box. In short, the measurements would be multiples of the shortest dimension. Then suppose we use this approach to measure a cereal box. That is, for the cereal box its shortest dimension is used to define the measurement instrument. Obviously, the box we are measuring affects our measurement instrument and our measurements would not possess the invariance property. Without invariance, our comparisons across different boxes would have limited utility.

The final issue we present brings us back to the classic definition of measurement mentioned above. Depending on which approach we use to measure anxiety (i.e., GSR, duration of headache, item responses, etc.), the measurements have certain inherent properties that affect how we interpret their information. For instance, the "duration of headaches" approach produces measurements that cannot be negative and that allow us to make comparative statements among people as well as to determine whether a person has a headache. These properties are a reflection of the fact that the measurements have not only a constant unit, but also a (absolute) zero point that reflects the absence of what is being measured. Invoking Stevens's (1946) levels of measurement taxonomy or Coombs's (1974) taxonomy, these numbers would reflect a *ratio* scale.

In contrast, if we use a GSR device for measuring anxiety, we would need to establish a baseline or a zero point by canceling out an individual's normal skin resistance static level before we measure the person's GSR. As a result, and unlike that of the ratio scale, this zero point is not an absolute zero but rather a relative one. However, all of our measurements would still have a constant unit and would be considered to be on an *interval* scale. Another approach to measuring anxiety is to ask an individual to rate his or her anxiety in terms of severity. This ratings approach would produce numbers that are on an *ordinal* scale. These approaches allow us to make comparative statements, such as "This person's anxiety level is greater than (or less than) that of another," or in the case of the interval scale, "This person's anxiety level is half as severe as that person's anxiety level." Alternatively, if our question simply requires the respondent to reply "yes," they are experiencing a symptom, or "no," they are not, then the "yes/no" responses would reflect a nominal scale. These various scenarios show that how we interpret and use our data needs to take into account the different types of information the observations carry.

In the following discussion, we present three approaches for establishing a correspondence between our observations and our latent variable. We begin by briefly introducing IRT, followed by classical test theory (CTT). Both of these approaches assume the latent variable is continuous. The last approach discussed, latent class analysis (LCA), is appropriate for categorical latent variables. Appendix F, "Mixture Models," addresses the situation when a latent variable is conceptualized as having categorical and continuous facets.

Item Response Theory

The term *theory* is used here in the sense that it is a paradigm that attempts to explain all the facts with which it can be confronted (Kuhn, 1970, p. 18). IRT is, in effect, a system of models that defines one way of establishing the correspondence between latent variables and their manifestations. It is not a theory in the traditional sense because it does not explain why a person provides a particular response to an item or how the person decides what to answer (cf. Falmagne, 1989). Instead, IRT is like the theory of statistical estimation. IRT uses latent characterizations of individuals and items as predictors of observed responses. Although some researchers (e.g., Embretson, 1984; Fischer & Formann, 1982) have attempted to use item characteristics to explain why an item is located at a particular point, for the most part, IRT like other scaling methods (e.g., Guttman Scalogram, Coombs Unfolding) treats the individual as a black box. (See Appendix E, “Linear Logistic Test Model [LLTM],” for a brief presentation of one of these explanatory approaches, as well as De Boeck and Wilson [2004] for alternative approaches.) The cognitive processes used by an individual to respond to an item are not modeled in the commonly used IRT models. In short, this approach is analogous to measuring the speed of an automobile without understanding how an automobile moves.¹

In IRT, persons and items are located on the same continuum. Most IRT models assume the latent variable is represented by a unidimensional continuum. In addition, for an item to have any utility, it must be able to differentiate among persons located at different points along a continuum. An item’s capacity to differentiate among persons reduces our uncertainty about their locations. This capacity to differentiate among people with different locations may be held constant or allowed to vary across an instrument’s items. Therefore, individuals are characterized in terms of their locations on the latent variable, and, at a minimum, items are characterized with respect to their locations and capacity to discriminate among persons. The gist of IRT is the (logistic or multinomial) regression of observed item responses on the persons’ locations on the latent variable and the item’s latent characterization(s).

Classical Test Theory

Like IRT, *classical test theory* (CTT) or *true score theory* also assumes the latent variable is continuous. CTT is the approach that most readers have been exposed to throughout their education. In contrast to IRT in which the item is the unit of focus, in CTT the respondent’s observed score on a whole instrument is the unit of focus. The individual’s observed score, X , is (typically) the unweighted sum of the person’s responses to an instrument’s items. In ability or achievement assessment, this sum reflects the number of correct responses.

CTT is based on the true score model. This model relates the individual’s observed score to his or her location on the latent variable. To understand this model, assume that an individual is administered an instrument an infinite independent number of times. On each of these administrations we calculate the individual’s observed score. The mean

of the infinite number of observed scores is the expectation of the observed scores (i.e., $\mu_i = \varepsilon(X_i)$). On any given administration of the instrument, the person's observed score will not exactly agree with the mean, μ , of the observed scores. This difference between the observed score and the mean is considered to be error. Symbolically, we may write the relationship between person i 's observed score, its expectation, and error as

$$X_i = \mu_i + E_i, \quad (1.1)$$

where E_i is the error score or the *error of measurement* (i.e., $E_i = X_i - \mu_i$); E_i is the capital Greek letter epsilon. Equation 1.1 is known as the *true score model*. In words, this model states that person i 's observed performance on an instrument is a function of his or her expected performance on the instrument plus error. Given that the error scores are considered to be random and $\mu_i = \varepsilon(X_i)$, then it follows that the mean error for an individual across the infinite number of independent administrations of the instrument is zero.

By convention, μ_i is typically represented by the Latin or Roman letter T . However, to be consistent with our use of Greek letters to symbolize parameters, we use the capital Greek letter tau, T . The symbol T represents the person's *true score* (i.e., $T_i = \varepsilon(X_i)$). The term *true score* should not be interpreted as indicating truth in any way. As such, *true score* is a misnomer. To avoid this possible misinterpretation, we refer to T_i (or μ_i) as individual i 's trait score.² A person's *trait score* represents his or her location on the latent variable of interest and is fixed for an individual and instrument. The common representation of the model in Equation 1.1 is

$$X_i = T_i + E_i. \quad (1.2)$$

Although Equation 1.1 may be considered more informative than Equation 1.2, we follow the convention of using T for the trait score.

There is a functional relationship between the IRT person latent trait (θ) and the CTT person trait characterization. This relationship is based on the assumption of parallel forms for an instrument. That is, each item has the same response function on all the forms. Following Lord and Novick (1968), assume that we administer an infinite number of independent parallel forms of an instrument to an individual. Then the *expected proportion of 1s* or *expected trait score*, $\varepsilon(T)$, across these parallel forms is equal to the average probability of a response of 1 on the instrument, given the person's latent trait and an IRT model. As a consequence, the IRT θ is the same as the expected proportion $\varepsilon(T)$ except for the difference in their scales of measurement. That is, θ has a range of $-\infty$ to ∞ , whereas for $\varepsilon(T)$ the range is 0 to 1. The expected trait score $\varepsilon(T)$ is related to the IRT latent trait by a monotonic increasing transformation. This transformation is discussed in Chapters 4 and 10.

In addition to the true score model, CTT is based on a set of assumptions. These assumptions are that, in the population, (1) the errors are uncorrelated with the trait scores for an instrument, (2) the errors on one instrument are uncorrelated with the

trait scores on a different instrument, and (3) the errors on one instrument are uncorrelated with the error scores on a different instrument. These assumptions are considered to be “weak” assumptions because they are likely to be met by the data. In contrast, IRT is based on “strong” assumptions.³ These IRT assumptions are discussed in the following chapter.

These CTT assumptions and the model given in Equation 1.1 (or Equation 1.2) form the basis of the psychometric concept of reliability and the validity coefficient. For example, the correlation of the observed scores on an instrument and the corresponding trait scores is the index of reliability for the instrument. Moreover, using the variances of the trait scores (σ_T^2) and observed scores (σ_X^2), we can obtain the population reliability of an instrument’s scores:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} \quad (1.3)$$

Because trait score variance is unknown, we can only estimate $\rho_{XX'}$. Some of the traditional approaches for estimating reliability are KR-20, KR-21, and coefficient alpha. An assessment of the variance of the errors of measurement in any set of observed scores may be obtained by substituting $\sigma_E^2 = \sigma_X^2 - \sigma_T^2$ into Equation 1.3 to get

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}) \quad (1.4)$$

The square root of σ_E^2 (i.e., σ_E) is referred to as the standard error of measurement. The *standard error of measurement* is the standard deviation of the errors of measurement associated with the observed scores for a particular group of respondents.

From the foregoing it should be clear that because an individual’s trait score is latent and unknown, then the error associated with an observed score is also unknown. Therefore, Equations 1.1 and 1.2 have two unknown quantities, an individual’s trait score and error score. Lord (1980) points out that the model given in Equations 1.1 or 1.2 cannot be disproved by any set of data. As a result, one difference between IRT and CTT is that with IRT we can engage in model–data fit analysis, whereas in CTT we do not examine model–data fit and simply assume the model to be true.

As may be obvious, the observed score X is influenced by the instrument’s characteristics. For example, assume a proficiency testing situation. An easy test administered an infinite number of independent times to an individual will yield a different value of T_i than a difficult test administered an infinite number of independent times to the same individual. This is analogous to the example of measuring the shoe and cereal boxes by using the shortest dimension of each box. In short, as is the case with the box example, in CTT person measurement is dependent on the instrument’s characteristics. Moreover, because the variance of the sample’s observed scores appears in both Equations 1.3 and 1.4, one may deduce that the heterogeneity (or lack thereof) of the observed scores affects both reliability and the standard error of measurement. Moreover, Equations 1.3 and 1.4 cannot be considered to solely be properties of the instrument, but rather also

reflect the sample's characteristics. In short, the instrument's characteristics affect the person scores, and sample characteristics affect the quantitative indices of the instrument (e.g., item difficulty and discrimination, reliability, etc.). Thus, Thurstone's (1928) idea of invariance does not exist in CTT. In contrast, with IRT it is possible to have invariance of both person and item characterizations. See Appendix G, "Dependency in Traditional Item Statistics and Observed Scores," for a demonstration of this lack of invariance with CTT. In addition, Gulliksen (1987) contains detailed information on CTT, Engelhard (1994, 2008) presents a historical view of invariance, and Holland and Hoskens (2003) show how CTT can be viewed as a first-order "approximation to a very general version" (p. 123) of IRT.

Latent Class Analysis

Unlike IRT's premise of a continuous latent variable, in *latent class analysis* (LCA) the latent variable is assumed to be categorical. That is, the latent variable consists of a set of mutually exclusive and exhaustive *latent classes*.⁴ To be more specific, "there exists a set of latent classes, such that the manifest relationship between any two or more items on a test can be accounted for by the existence of these basic classes and by these alone" (Stouffer, 1950, p. 6). In LCA, the comparison of individuals involves comparing their latent class memberships rather than their locations on a continuous latent variable.

For an understanding of the nature of a categorical latent variable, we turn to two empirical studies. The first is a study of the nosologic structure of psychotic illness by Kendler, Karkowski, and Walsh (1998). In their study, these authors conceptualized this latent variable as categorical. Their LCA showed that their participants belonged to one of six classes: (1) classic schizophrenia, (2) major depression, (3) schizophreniform disorder, (4) bipolar-schizomania, (5) schizodepression, and (6) hebephrenia. The second study involves cheating on academic examinations (Dayton & Scheers, 1997); the latent variable is cheating. The LCA of the investigators' data revealed a structure with two latent classes. One class consisted of persons who were "persistent cheaters," whereas the second class consisted of individuals who would either exhibit "opportunistic" cheating or might not cheat at all.

In both of these examples, we can see that respondents differ from one another on the latent variable in terms of their class membership rather than in terms of their locations on a continuum. In Appendix F, "Mixture Models," we discuss an approach in which we combine LCA and IRT. That is, we can conceptualize academic performance as involving both latent classes and continua. For example, we can have one class of persistent cheaters and another class of noncheaters. Within each class there is a proficiency variable continuum. Therefore, the cheater class has its own continuum on which we can compare individual performances. Similarly, the noncheater class has a separate continuum that we use to compare the noncheaters' performances with one another. These latter comparisons are not contaminated by the cheaters' performances, and the noncheaters are not disadvantaged by the presence of the cheaters.

In general, LCA determines the number of latent classes that best explains the data

(i.e., determining the latent class structure). This process involves comparing models that vary in their respective number of classes (e.g., a one-class model and a two-class model). Determining the latent class structure involves not only statistical tests of fit, but also the interpretability of the solution. With each class structure, one has estimates of the items' characteristics. Based on these item characteristics and the individuals' responses, the respondents are assigned to one of the latent classes. Subsequent to this assignment, we obtain estimates of the relative size of each class. These relative sizes are known as the latent class proportions (π_v , where v is the latent class index). The sum of the latent class proportions across the latent classes is constrained to 1.0. For example, if the latent variable, say algebra proficiency, has a two-class structure, then π_1 might equal 0.65 and $\pi_2 = 1 - 0.65 = 0.35$. Moreover, our latent classes' interpretation may reveal that the larger class (i.e., $\pi_1 = 0.65$) consists of persons who have mastered algebraic problems, whereas the other class consists of individuals who have not mastered the problems. In short, the data's latent structure consists of masters and nonmasters.

One may conceive of a situation in which if one had a large number of latent classes and if they were ordered, then there would be little difference between conceptualizing the latent variable as continuous or as categorical. In point of fact, a latent class model with a sufficient number of latent classes is equivalent to an IRT model. For example, for a data set with four items, then a latent class model with at least three latent classes would provide "equivalent" item characterizations as an IRT model that uses only item location parameters. Appendix F, "Mixture Models," contains additional information about LCA.

Summary

Typically, measurement is viewed as analogous to using a ruler to measure the length of an object. In effect, this is analogous to Stevens's (1946) definition of measurement in that the ruler provides the "rules" and the numeric values associated with the ruler's tick marks provide the numeric labels. However, Stevens's definition invites misinterpretation. Although one can infer from his definition that he is describing an act or a process, this aspect of the definition is not made salient. Moreover, by focusing only on the assignment of numbers, one is left with the impression that measurement results in only a set of numeric labels. We consider measurement to be a process by which one attempts to understand the nature of a variable by applying mathematical techniques. The result may or may not be numeric labels and may or may not involve a continuous variable. For example, LCA is a measurement paradigm that allows one to understand the nature of a latent variable, such as ethnocentrism or test anxiety, without resulting in numeric labels. The use of LCAs involves the application of mathematical techniques that results in individuals being classified into latent classes and an assessment of how well the class structure describes the manifest data.

The term *manifest data* refers to the information obtained by direct observation, whereas the term *latent* refers to the information obtained on the basis of additional assumptions and/or by making inferences from the original (manifest) data (Lazarsfeld,

1950). Presumably, one or more latent variables can account for the patterns or relationships that are evident in the manifest data. Therefore, a *manifest variable* is an observed manifestation of one or more *latent* (i.e., unobservable) *variables*. Above we outlined different paradigms that allow the tools of mathematics to be applied to explaining manifest observations from a latent variable perspective.

A latent variable may be conceptualized as continuous, categorical, or some combination of the two. When the variable is conceptualized as continuous, then the use of CTT or IRT may be the appropriate mathematical technique. However, if the variable is conceptualized as categorical, then LCA may be the most appropriate psychometric method to use. It is possible to conceptualize the latent space as a set of latent classes, within each of which there is a continuum, or as a combination of latent classes and a latent continuum. In this situation, a mixture of IRT and LCA may be considered an appropriate representation of the latent space.

As part of measurement, it is necessary to operationalize the variable(s) of interest (i.e., provide operational definitions). The measurement process also involves assessing how much information the measures yield about the participants (e.g., reliability) as well as how well the measures reflect the latent variable(s) (i.e., validity).

When IRT is appropriate and when there is model–data fit, then IRT offers advantages over CTT. For instance, with IRT, our person location estimates are invariant with respect to the instrument, the precision of these estimates is known at the individual level and not just at the group level (as is the case with Equation 1.4), and the item parameter estimates transcend the particular sample used in their estimation. Moreover, unlike CTT, with IRT we are able to make predictive statements about respondents' performance as well as examine the tenability of the model vis-à-vis the data. In the next chapter, the simplest of the IRT models is presented. This model, the Rasch or one-parameter logistic model, contains a single parameter that characterizes the item's location on the latent variable continuum. We show how this single-item parameter can be used to estimate a respondent's location on a latent variable.

Notes

1. Although understanding how the automobile moves is not necessary to measure the speed with which it moves, nonetheless fully understanding how the automobile moves can lead to an improved measurement process.
2. A trait is exemplified primarily in the things that a person can do (Thurstone, 1947) and is any “distinguishable, relatively enduring way in which one individual varies from another” (Guilford, 1959, p. 6). However, we do not consider traits to be rigidly fixed or predetermined (see Anastasi, 1983).
3. There is an implicit unidimensionality assumption in CTT. That is, for observed scores to have any meaning, they need to represent the sum of responses to items that measure the same thing. For instance, assume that an examination consists of five spelling questions and five single-digit addition problems. Presumably, our examination data would consist of two dimensions representing spelling and addi-

tion proficiencies. If a person had an observed score of 5, it would not be possible to determine whether they are perfect in spelling, perfect in addition, or in some combination of spelling and addition proficiencies. In this case, the observed score has no intrinsic meaning. In contrast, if the examination consists of only spelling questions, then the score would indicate how well a person could spell the questions on the test and would have intrinsic meaning.

4. Both IRT and LCA can be considered to be special instances of the general theoretical framework for modeling categorical variables, known as latent structure analysis (LSA; Lazarsfeld, 1950). Moreover, both linear and nonlinear factor analysis may be regarded as special cases of LSA (McDonald, 1967).

2

The One-Parameter Model

IRT may be applied to many possible construct domains. These constructs involve psychological constructs, such as neuroticism, motivation, social anxiety, cognitive development, consumer preferences, and proficiency. Although each of these latent variables could be conceptualized as being categorical (e.g., consisting of only latent classes), in the current context we would conceptualize each of them as a continuum. Whatever the construct of interest may be, we assume that it is manifested through an individual's responses to a series of items. The simplest IRT model that could be used in this situation is one that characterizes each item in terms of a single parameter. This parameter is the item's location on the latent continuum that represents the construct. The concept that an item has a location is not new and may be traced back Thurstone (1925, 1928); also see Andrich (1978a), Lumsden (1978), and Yen (1986). In this chapter, a one-item parameter IRT model is conceptually developed. In the context of this model, the general principles and assumptions underlying IRT, as well as a parameter estimation approach, are presented. More sophisticated estimation approaches and more complicated models are discussed in subsequent chapters.

Conceptual Development of the Rasch Model

Assume that we are interested in measuring the mathematics proficiencies of a group of individuals. Although we cannot directly observe mathematics proficiency, we can infer its existence through behavioral observations. As such, mathematics proficiency is considered to be a latent variable, and, in the current context, this means that it is conceptualized as a latent continuum. To assess the individuals' mathematics proficiencies, they are administered an instrument containing five questions. Their responses to this instrument constitute our behavioral observations. The instrument's items are located at various points along the continuum representing mathematics proficiency. For instance, Figure 2.1 depicts the locations of the items on the continuum, as well as one person's location on the same continuum. Typically, we use standard score-like values (i.e., z -scores) to mark off the continuum and represent the metric in IRT.

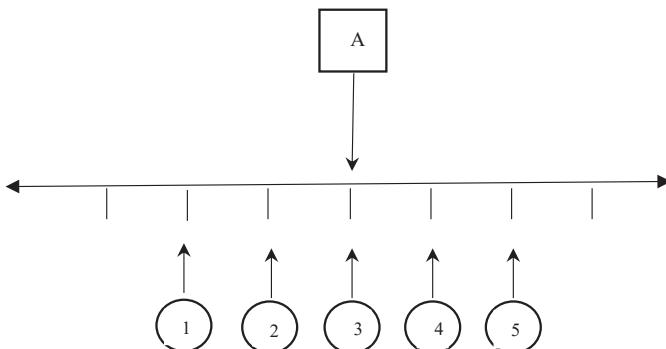


FIGURE 2.1. Graphical representation of latent variable continuum with five items (circles).

For this continuum, assume the upper end of the continuum indicates greater mathematics proficiency than does the lower end. This means that items located toward the right side require an individual to have greater proficiency to correctly answer the items than items located toward the left side. As can be seen, our instrument's items are located throughout the continuum with some above 0 and others below 0. For instance, the first item is located at -2 , the second item at -1 , and so on. We use the Greek letter δ (delta) to represent an item's location and δ_j to represent the j th item location on this continuum. Using this notation, we see that the first item's location is represented as $\delta_1 = -2$ and the fifth's as $\delta_5 = 2$. Moreover, the Greek letter θ (theta) is used to represent the person location on this continuum. In the context of this example, a person's location reflects their mathematics proficiency. According to the figure, person A is located at 0.0 (i.e., $\theta_A = 0.0$). As should be clear from Figure 2.1, both items and persons are located on the same continuum.

One implication of locating both persons and items on the same continuum is that it is possible to make comparative statements about how a typical person might respond to an item. For example, because the lower end of the continuum represents less mathematics proficiency than the upper end, items that are located in the lower end require less proficiency to be correctly answered than those in the upper end. As a result, chances are that a person located at 0 will correctly respond to items located in the lower end of the continuum (e.g., item 1 with a $\delta_1 = -2$). However, if we administer an item located closer to 0, say item 2 with $\delta_2 = -1$, then chances are that the person will respond correctly, but we recognize that there is an increased possibility that they may respond incorrectly. This incorrect response may be due to a lapse in being able to recall relevant information, the tip-of-tongue phenomenon, or another such cause. Similarly, administering an item, such as item 4 ($\delta_4 = 1$), to a person located at 0 will likely result in an incorrect response, but there is still a sizeable chance that they may correctly answer the item because of the closeness in the proficiency required by the item and that which the person possesses. In other words, the greater the distance between the person and item locations, the greater the certainty we have about how the person is expected to respond to the item. However, as this distance approaches zero, then the more likely we

are to say that there is a 50–50 chance that the person will correctly respond to the item. These expectations about how a person will respond are expressed probabilistically (i.e., “chances of a correct response are . . . ”).

Although the foregoing may make intuitive sense, one might ask, “Are there data that support a pattern of an increasing probability of a correct response as person location increases?” In Figure 2.2 we see that the answer is yes. This figure shows that the proportion of individuals correctly responding to an item is an S-shape (sigmoidal) function of their standard scores on a test. In this case, the participants were administered an examination to assess their latent mathematics proficiency. The complete data from this administration are presented in Table 2.1.

From the graph we see that as the z -scores increase, there is an increase in the proportion of examinees correctly responding to the item; however, this increase is not constant across the continuum. Moreover, we see that as one progresses beyond a z of 1, the points begin to form a plateau. Conversely, as one progresses below a z of -1, the proportions also start to level off. However, there is a range on the z -score metric around -0.5 where the proportion of individuals with a correct response is around 0.50. That is, this is the point at which we would say that there is a 50–50 chance that the person will correctly respond to the item.

We can trace the nonlinear pattern of the empirical proportions in Figure 2.2 and obtain an empirical *trace line* (Lazarsfeld, 1950). This trace line would clearly be a sigmoid or S-shaped curve (i.e., an ogive). However, rather than being satisfied to simply

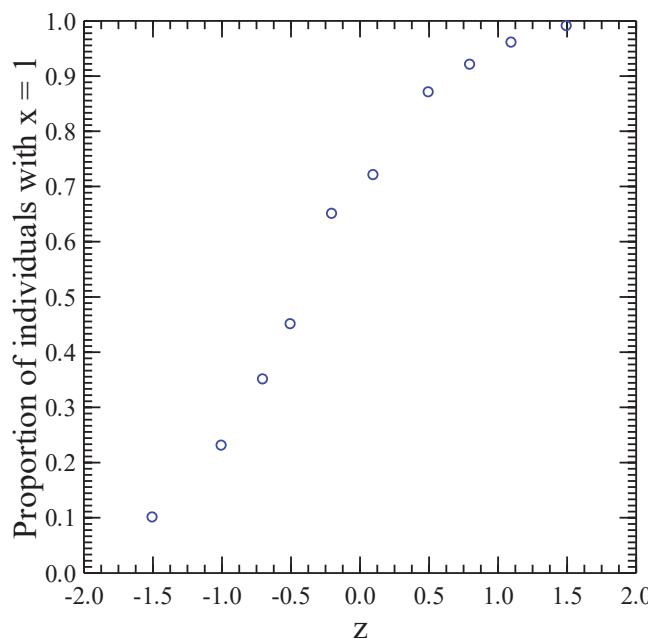


FIGURE 2.2. Empirical proportions of individuals with a correct response ($x = 1$) on one item as a function of standardized number correct scores.

TABLE 2.1. Response Patterns and Their Frequency on a Mathematics Test

Pattern	Frequency	X
00000	691	0
10000	2280	1
01000	242	1
00100	235	1
00010	158	1
00001	184	1
11000	1685	2
10100	1053	2
01100	134	2
10010	462	2
01010	92	2
00110	65	2
10001	571	2
01001	79	2
00101	87	2
00011	41	2
11100	1682	3
11010	702	3
10110	370	3
01110	63	3
11001	626	3
10101	412	3
10011	166	3
01101	52	3
01011	28	3
00111	15	3
11110	2095	4
11101	1219	4
11011	500	4
10111	187	4
01111	40	4
11111	3385	5

Note. N = 19,601.

describe the pattern, we can develop a model that incorporates our ideas about how an observed response is governed by a person's location in order to be able to predict response behavior. The nonlinearity shown in Figure 2.2 eliminates using the linear regression of proportions on the person locations to predict response behavior. Because this ogival pattern is evident in cumulative distributions, such as the cumulative normal distribution or the logistic distribution, we might consider using one of these for our modeling. In point of fact, we use the logistic function in the following because of its simplicity. (Use of the cumulative normal function is discussed in Appendix C.)

In its simplest form, the logistic model may be presented as

$$p(x) = \frac{e^z}{1+e^z} \quad (2.1)$$

where $p(x)$ is the probability of value of 1 when the predictor takes on the value of x , e is a constant equal to $2.7183 \dots$, and z is some linear combination of, for example, predictor variable(s) and a constant. By appropriately specifying z , we can arrive at a model for predicting response behavior on an item j .

To specify z , we return to the idea above that the distance between the person and the item locations (i.e., $(\theta - \delta_j)$) is an important determinant of the probability of their response (cf. Rasch, 1980; Wright & Stone, 1979). Therefore, letting $z = (\theta - \delta_j)$ results in a model that would allow one to predict the probability of a response of 1 as a function of both the item and person locations.¹ By substitution of this z into the logistic model, we have

$$p(x_j = 1 | \theta, \delta_j) = \frac{e^{(\theta - \delta_j)}}{1 + e^{(\theta - \delta_j)}}, \quad (2.2)$$

where $p(x_j = 1 | \theta, \delta_j)$ is the probability of the response of 1 (i.e., $x_j = 1$), θ is the person location, and δ_j is item j 's location. This model is called the *Rasch* model (Rasch, 1980).² Expressed in words, Equation 2.2 says that the probability of a response of 1 on item j is a function of the distance between a person located at θ and the item located at δ_j . (Technically, we are talking about the probability of a randomly selected individual from those located at θ .) The right side maps the (potentially infinite) distance between the person's location and the item's location onto the $[0, 1]$ probability scale. A response of 1 simply indicates that an event occurred or we observed a success. (We use the phrase "response of 1" instead of "correct response" because the instrument may not be an examination; given a proficiency context, we may refer to the response as correct or incorrect.) For convenience, p_j is used for $p(x_j = 1 | \theta, \delta_j)$ in the following.

The theoretical range of the item locations δ s, as well as the person locations θ s, is from $-\infty$ to ∞ . However, typical item and person locations fall within -3 to 3 . In proficiency testing, the item locations are referred to as *item difficulties*.³ In general, items located somewhat below 0.0 are said to be "easy" items (e.g., below -2.0), and items somewhat above 0.0 are "hard" items (e.g., above 2.0). In general, the items that are considered to be "easy" are the ones that persons with low proficiencies have a tendency to answer correctly. Conversely, the "harder" items are the items that persons with high proficiencies tend to get correct. Items around 0.0 are considered to be of "average difficulty."

As an example of using the Rasch model to predict response behavior, assume that we administer a mathematics item located at 1 (i.e., $\delta = 1$) to individuals located at 0 (i.e., $\theta = 0$). According to the model in Equation 2.2, the probability of a correct response for a randomly selected individual from this group would be

$$p_j = \frac{e^{(0-1)}}{1 + e^{(0-1)}} = 0.2689$$

That is, the probability that a randomly selected person located at 0 will correctly respond to this item is only 0.2689. The magnitude of this probability should not be surprising, given that this item is located above (or to the right of) the individual's location,

and so the item requires more mathematical proficiency to be correctly answered than the person possesses. Actually, chances are that a person located at 0 will *incorrectly* respond to this item rather than respond correctly because the probability of an incorrect response to this item by someone located at 0 is $1 - 0.2689 = 0.7311$.

Another way of interpreting our probabilities is to convert them to the *odds* of a correct response on the item. For example, converting these probabilities to odds, we find that the odds of a response of 1 are approximately 1 to 2.7, or that it's almost three times more likely that the person will *incorrectly* respond to the item than correctly respond. Appendix G, "Odds, Odds Ratios, and Logits," contains more information on odds.

For a given item location, the substitution of different values of θ into Equation 2.2 produces a series of probabilities that when graphed show a pattern similar to that shown in Figure 2.2. The line produced by the model given in Equation 2.2 is referred to as an *item characteristic curve* (ICC; Lord, 1952), an item curve (Tucker, 1946), *item operating characteristic* (Green, 1954), *item characteristic function* (Lord & Novick, 1968), or a trace line (Lazarsfeld, 1950). We will call it an *item response function* (IRF). Example IRFs are shown in Figure 2.3 and are discussed below. For the Rasch model, the item's location, δ , is defined at the point of inflection or the "middle" point of the IRF; an inflection point is where the slope of a function changes direction and sign. Because the Rasch model IRF has a lower asymptote of 0 and an upper asymptote of 1, this midpoint has a value of 0.50. Therefore, for the Rasch model the item's location corresponds to the point on the continuum where the probability of a response of 1 is 0.50.

The One-Parameter Model

In the IRT literature, one sometimes sees reference to a one-parameter model. In this section we present the one-parameter model, and in the subsequent section we discuss whether the Rasch model and the one-parameter model should be considered distinct models.

Figure 2.3 shows a series of IRFs overlaid on the item data shown in Figure 2.2. Each of these IRFs uses an estimated item location of -0.37 (i.e., $\hat{\delta}_j = -0.37$).⁴ The dashed IRF (labeled Rasch: long dash) is based on the Rasch model and is created by substituting $\hat{\delta} = -0.37$ for δ in Equation 2.2 and using values of θ from -2.0 to 2.0 .

As we see, the predicted Rasch IRF is not as steep as the observed response function (i.e., the empirical trace line suggested by the proportions). To better match the empirical trace line, we need to increase the slope of the IRF. To do this we revisit the exponent in Equation 2.2. This exponent, $(\theta - \hat{\delta}_j)$, can be considered to have a multiplier whose value is 1. That is, if we symbolize this multiplier by α , then the exponent becomes $\alpha(\theta - \hat{\delta}_j)$. Distributing α across $(\theta - \hat{\delta}_j)$ and by letting

$$\gamma_j = -\alpha\delta_j \quad (2.3)$$

our exponent becomes

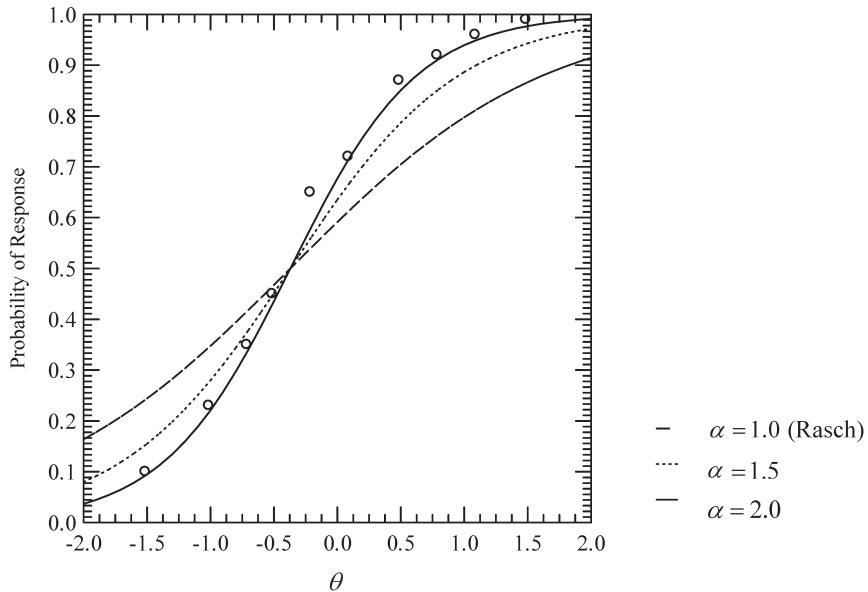


FIGURE 2.3. Empirical proportions and IRFs corresponding to different discrimination values.

$$\alpha(\theta - \delta_j) = \alpha\theta - \alpha\delta_j = \alpha\theta + \gamma_j. \quad (2.4)$$

Equation 2.4 is the slope–intercept parameterization of the exponent. In this form we have the equation for a straight line, where α represents the slope and γ_j symbolizes the intercept; “ $\alpha\theta + \gamma_j$ ” is sometimes referred to as being in linear form. Although the intercept is related to the item location, it is not the item’s location. To obtain the item’s location we would use

$$\delta_j = -\frac{\gamma_j}{\alpha} \quad (2.5)$$

To better understand the slope–intercept form, examine Figure 2.4. (We refer to the line in the graph as a *logit regression line*.) This figure shows $\alpha\theta + \gamma_j$ as a function of θ for an item located at -0.37 . As can be seen, the line’s intercept (γ_j) equals 0.37 and the slope (α) is 1 . From these values we can obtain the item’s location on the θ continuum using Equation 2.5

$$\delta_j = -\frac{\gamma_j}{\alpha} = -\frac{0.37}{1.0} = -0.37 \quad (2.6)$$

Because α is directly related to the logit regression line’s slope, a change in α leads to a change in the line’s slope. The effect of changing α “passes through” the reparameterization of the slope–intercept form into the $\alpha(\theta - \delta_j)$ deviate form. That is, the slope of the IRF may be modified by changing the value of α .⁵ For example, by increasing α we arrive at the other two IRFs shown in Figure 2.3. The solid and dotted IRFs are the IRFs when $\alpha = 1.5$ or $\alpha = 2.0$, respectively. As we see, using an $\alpha = 2.0$ results in a predicted

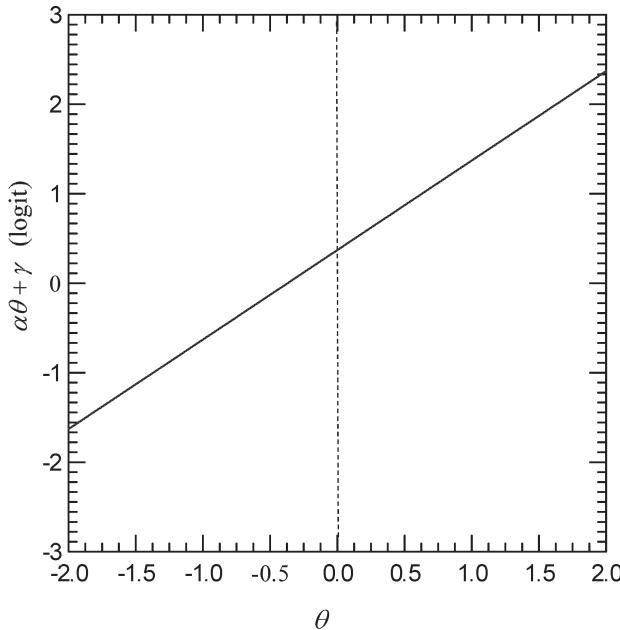


FIGURE 2.4. Logit space plot for item located at -0.37 .

IRF that almost perfectly matches the empirical trace line. Stated another way, in this case the net effect of increasing the value of α is to improve the fit of the model to the data.

We can rewrite Equation 2.2 to explicitly incorporate α . Doing this produces the one-parameter logistic (1PL) model

$$p(x_j = 1 | \theta, \alpha, \delta_j) = \frac{e^{\alpha(\theta - \delta_j)}}{1 + e^{\alpha(\theta - \delta_j)}} \quad (2.7)$$

For computational efficiency Equation 2.7 is sometimes written as

$$p(x_j = 1 | \theta, \alpha, \delta_j) = \frac{1}{1 + e^{-\alpha(\theta - \delta_j)}} \quad (2.8)$$

The lack of a subscript on α means that α does not vary across items. As such, the corresponding IRFs do not cross one another.⁶ Sometimes “ $\alpha(\theta - \delta_j)$ ” is referred to as a *logistic deviate*. For simplicity of presentation p_j is used in lieu of $p(x_j = 1 | \theta, \alpha, \delta_j)$ in the following.

Because α is related to the IRF’s slope, it reflects how well an item discriminates among individuals located at different points along the continuum. As a consequence, α is known as the *item discrimination parameter*. As an aid to understanding, this assumes we have three items with different α s, but located at 0.0 (i.e., $\delta_1 = \delta_2 = \delta_3 = 0$). Our three discrimination parameters are 0 , 1 , and 2 . In addition, we have a respondent A located at -1 ($\theta_A = -1$) and another respondent B located at 1 (i.e., $\theta_B = 1$).

For the item with $\alpha = 0.0$ our IRF, as well as logit regression line, is horizontal. As a result, the predicted probabilities of a response of 1 for the two respondents is 0.5. In this case, the item does not provide any information for differentiating between the two respondents. This lack of discriminatory power is a direct function of $\alpha = 0.0$. In contrast, with the second item ($\alpha = 1$) we have different predictions for our respondents; for respondent A the $p_2 = 0.2689$ and for person B the $p_2 = 0.7311$. Therefore, this item's α allows us to distinguish between the two respondents.

Developing this idea further, we find that the third item ($\alpha = 2.0$) would have the steepest IRF (and logit regression line) of the three items. This steepness is reflected in a greater difference in the predicted probabilities for our respondents than is seen with the previous two items. That is, for this item respondent A has a $p_3 = 0.1192$ and for person B we have $p_3 = 0.8808$. In short, the magnitude of the difference in these predicted probabilities is a direct function of the item's α . Therefore, items with larger α s (i.e., steeper logit regression lines and IRFs) do a better job of discriminating among respondents located at different points on the continuum than do items with smaller α s.

The One-Parameter Logistic Model and the Rasch Model

To summarize, both the 1PL and Rasch models require that items have a constant α , but they allow the items to differ in their locations. For the Rasch model this constant is 1.0, whereas for the 1PL model the constant α does not have to be equal to 1.0. Mathematically, the 1PL and the Rasch models are equivalent. The values from one model can be transformed into the other by appropriate rescaling. Use of the Rasch model sets α to 1.0, and this constant value is absorbed into the metric used in defining the continuum; this is demonstrated in Chapter 4.

However, to some the Rasch model represents a different *philosophical* perspective than that embodied in the 1PL model. The 1PL model is focused on fitting the data as well as possible, given the model's constraints. In contrast, the Rasch model is a model used to *construct* the variable of interest (cf. Andrich, 1988; Wilson, 2005; Wright, 1984; Wright & Masters, 1982; Wright & Stone, 1979). In short, this perspective says that the Rasch model is the standard by which one can create an instrument for measuring a variable. This perspective is similar to that seen in Guttman Scaling and Coombs Unfolding (Coombs, 1950) and is analogous to what is done in the physical sciences.⁷ For example, consider the measurement of time. The measurement of time involves a repetitive process that marks off equal increments (i.e., units) of the (latent) variable time. To measure time, we need to define our unit (e.g., a standard period of oscillation). With the Rasch model the unit is defined as the logit. That is, the unit is the distance on our continuum that leads to an increase in the odds of success by a factor equal to the transcendental constant e . Therefore, analogous to time measurement, our measurements with a one-parameter model are based on the (repetitive) use of a unit that remains constant across our metric.

For simplicity in the following discussion we use the general term 1PL model to refer to both $\alpha = 1.0$ (i.e., the Rasch model) and the situation where α is equal to some

other constant. However, when we use the term *Rasch model*, we are referring to the situation when $\alpha = 1.0$ and a measurement philosophy that states that the Rasch model is the basis for constructing the variable of interest.

Assumptions Underlying the Model

IRT models assume that response data are a manifestation of one or more person-oriented latent dimensions or factors. This is typically referred to as the *dimensionality assumption* and is reflected both in the models and in their graphical representations. For instance, in the 1PL model we use a single person location variable, θ , to reflect that one latent variable accounts for a person's response behavior. Moreover, in the 1PL model's conceptual development, as well as in its IRF, this assumption is reflected in a single continuum to represent the latent variable (see Figures 2.1 and 2.3). All the models presented in Chapters 2–9 assume a single latent person variable. Therefore, for these models the dimensionality assumption is referred to as the *unidimensionality assumption*. Specifically, the *unidimensionality assumption* states that observations on the manifest variables (e.g., the items) are solely a function of a single continuous latent person variable. If one has a unidimensional latent space, then the persons may be located and compared on this latent variable. In terms of our example, this assumption states that there is a single latent mathematics proficiency variable that underlies the respondents' performance on our instrument. In contrast, if we needed to also know the respondents' locations on an obsessiveness latent variable to account for their performance on our instrument, then the response data would be best modeled using a two-dimensional, not a unidimensional, model.

We view the unidimensionality assumption as representing an ideal situation analogous to the homogeneity of variance assumption in analysis of variance (ANOVA). In practice, there will most likely be some degree of violation of the unidimensionality assumption. This degree of violation may or may not be problematic. That is, although the data may in truth be a manifestation of, for example, two latent variables, a unidimensional model may provide a sufficiently accurate representation to still be useful. (This is similar to an ANOVA in which the homogeneity of variance assumption is violated, but the F test is still useful under certain conditions.) Of course, in some situations the degree of violation may be so large that a unidimensional model is not useful. In these situations, one might consider the use of a multidimensional model (Chapter 10) or some other approach to modeling the data. However, regardless of whether or not one uses a unidimensional model, it should be noted that whether the *estimated* θ s are meaningful and useful is a validity issue. In short, the estimated θ s, in and of themselves, do not guarantee that the latent variable that is intended to be measured (e.g., mathematics proficiency) is, in fact, measured.

A second assumption is that the responses to an item are independent of the responses to any other item conditional on the person's location. This assumption is referred to as *conditional independence* or *local independence*. We use the term *conditional*

independence in the following because we consider it to be more descriptive than the term *local independence*.

In the unidimensional case, the conditional independence assumption says that how a person responds to a question is determined solely by their location on the latent continuum and not by how they respond to any other question on the examination. If this were not true, then more than the person's, say, mathematics ability, would be affecting their responses and one would have a nonunidimensional situation. Given this interpretation, it is not surprising that sometimes the unidimensionality assumption and the conditional independence assumption are discussed as being one and the same. However, they may not be one and the same in all cases; also see Goldstein (1980). For instance, certain instrument formats lead to a dependency among item responses that does not appear to invoke additional *latent* variables to define the latent space. For example, we might have a series of questions that all relate to the same passage or a set of hierarchically related items in which answering later items is based, in part, on answer(s) to earlier item(s). With these formats, item responses will most likely violate the conditional independence assumption.

In contrast, there are cases of item interdependency that are due to additional latent variables. For example, consider the case of *speededness* in which an individual has insufficient time to respond to all the items on an instrument. In this situation, unless we use an additional latent variable, such as an individual's rapidity in defining the latent space, then conditional independence is violated for the speeded items. That is, the unidimensionality assumption is violated because we have two latent person variables, rapidity and the target latent variable (e.g., mathematics proficiency). Furthermore, to the extent that the latent variable rapidity is associated with gender or ethnic group differences, then one may also observe that the speeded items exhibit differential item functioning (Chapter 12). Verhelst, Verstralen, and Jansen (1997) and Roskam (1997) both present models for time-limited measurement.

Strictly speaking, the conditional independence assumption states that for "any group of examinees all characterized by the same values $\theta_1, \theta_2, \dots, \theta_k$, the (conditional) distributions of the item scores are all independent of each other" (Lord & Novick, 1968, p. 361).⁸ That is, when all the latent variables that define the complete latent space are known and taken into account, then the item responses are independent of one another. Therefore, the conditional independence assumption applies not only to unidimensional, but also to the multidimensional IRT models.

A third assumption is the *functional form* assumption. This assumption states that the data follow the function specified by the model. For instance, for Equations 2.2 and 2.7 the functional form is that of an "S"-shaped curve. This ogival form matches the empirical data for the item shown in Figure 2.2. (Although these data were modeled using a logistic function, an alternative approach would be to use a probit strategy; see Appendix C.)

In the context of the 1PL model, the functional form assumption also embodies the notion that all items on an instrument have IRFs with a common lower asymptote of 0 and a common slope. This common slope (i.e., a constant α) across items is reflected in parallel IRFs. As is the case with the unidimensionality assumption, this assumption is rarely exactly met in practice. However, if the IRFs are parallel within sampling error,

then this is interpreted as indicating model–data fit. Several different ways of determining model–data fit are addressed in the following chapters.

An Empirical Data Set: The Mathematics Data Set

To demonstrate the principles underlying person and item parameter estimation, we use response data from the administration of a five-item mathematics examination. Consistent with the measurement issues discussed in Chapter 1, we assume that we have content validity evidence for our instrument. Although there is some controversy concerning the concept of content validity, in this book we assume that it is a useful concept. See Sireci (1998) for a discussion of the concept of content validity. We conceptualize mathematics proficiency as a continuous latent variable.

Although our example data come from proficiency assessment, we could have just as easily used data from a personality, attitude, or interest inventory. Moreover, our IRT model does not make an assumption about the item response format used on our instrument. Whether the questions, for example, use a multiple-choice, open-ended, true-false, forced-choice, or fill-in-the-blank response format is irrelevant. All that matters is that the data analyzed are dichotomous and that the assumptions are tenable. The appropriateness of the model to the data is a fit issue.

For our example, the dichotomous data were determined by classifying the examinees' responses into one of two categories, correct or incorrect. If the examinee correctly performed the mathematical operation(s) on an item, then their response is categorized in the “correct” category and assigned a value of 1. Otherwise, their response is categorized as incorrect and received a value of 0. Table 2.1 contains the response patterns and their frequency of occurrence.

With binary data and five items there are $2^L = 2^5 = 32$ possible unique response patterns, where L is the number of items. As can be seen from Table 2.1, each of these possible patterns is observed in the data set. There are 691 persons who did not correctly respond to any of the items (i.e., the response pattern 00000 with an $X = 0$ or a “zero score”), and there are 3385 persons who obtained a perfect score of 5 (i.e., $X = 5$, for the response pattern 11111). For pedagogical purposes, the items are presented in ascending order of item location. Therefore, one would expect that if a person had item 2 correct, then that person should have also had item 1 correct.⁹

Conceptually Estimating an Individual's Location

In practice, we do not know either the item parameters (e.g., the δ s) or the person parameters (θ s). Because, in general, our interest is in estimating person locations, we begin with estimating respondent locations on the latent continuum. For estimating a respondent's location ($\hat{\theta}$), we assume that we know the items' locations on the latent continuum. Although we conceptually present the estimation process here, in Appendix A we furnish a mathematical treatment and address obtaining the item location estimates in Appendix B.

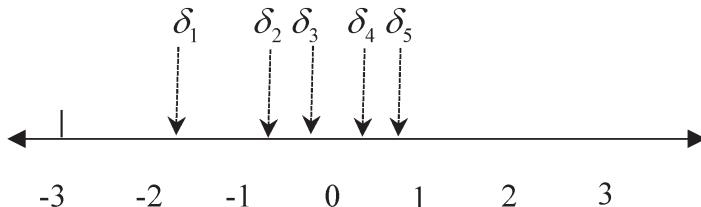


FIGURE 2.5. Graphical representation of the location of items on the five-item mathematics test.

Figure 2.5 shows the locations of our mathematics items. Item 1 has a location at -1.9 (i.e., $\delta_1 = -1.9$), item 2 is located at $\delta_2 = -0.6$, and the remaining items are located at $\delta_3 = -0.25$, $\delta_4 = 0.30$, and $\delta_5 = 0.45$; these values correspond to their locations in Figure 2.5. Stated in words, item 1 is almost two units below the zero point, and item 5 is almost a half unit above the zero point. The units defining the metric are called *logits*; see Appendix G, “Odds, Odds Ratios, and Logits,” for information on logits. In general, and given that we are assessing mathematics proficiency, item 1 would tend to be considered an “easy” item and is comparatively easier than the remaining items on the instrument. Items 2 through 5 are generally considered to be of “average” difficulty.

To demonstrate person location estimation, we arbitrarily select the response pattern of 11000 (i.e., two correct responses followed by three incorrect responses, $X = 2$). In this case, the estimation of a person’s location is conceptually equivalent to asking the question, “Which θ has the highest likelihood of producing the pattern 11000?” To answer this question, we need to perform a series of steps. The first step is to calculate the probability of each response in the pattern according to Equation 2.2. The second step is to determine the probability of the response pattern. Step 2 is accomplished by capitalizing on the conditional independence assumption (i.e., for a given θ the responses are independent of one another). This assumption allows us to apply the multiplication rule for independent events to the item probabilities to obtain the probability for the pattern given θ . The third “step” is to re-perform steps 1 and 2 for a range of θ values. For our example, the range of θ will be from -3 to 3 . The final step is to determine which of the various values of θ from step 3 has the highest likelihood of producing the pattern 11000.

Equation 2.2 specifies the probability of a response of 1, and its complement indicates the probability of a response of 0 (i.e., $p(x_j = 0) = 1.0 - p(x_j = 1)$). For the pattern in question (11000), items 1 and 2 are correctly answered and items 3–5 are incorrectly answered. Therefore, given our range of interest, the probability of a correct response to item 1 by someone located at -3.0 is

$$p(x_1 = 1 | \theta = -3.0, \delta_1 = -1.9) = 0.2497.$$

For our second item, and using δ_2 in lieu of δ_1 , the probability of a correct response to item 2 for someone located at -3.0 is

$$p(x_1 = 1 | \theta = -3.0, \delta_2 = -0.6) = 0.0832.$$

These probabilities reflect what one would expect—namely, that a randomly selected person located at -3.0 has a higher probability of correctly answering the easiest item on the instrument than of correctly answering a comparatively harder item.

The responses to items 3 through 5 are incorrect. Therefore, to determine the probability of an incorrect response, we use the complement of Equation 2.2 for items 3 through 5. Item 5 is used to demonstrate obtaining the probability of an incorrect response. For item 5 the probability of an incorrect response for someone located at -3.0 is

$$p(x_5 = 0 | \theta = -3.0, \delta_5 = 0.45) = 1.0 - p(x_5 = 1 | \theta = -3.0, \delta_5 = 0.45) = 1.0 - 0.0308 = 0.9692$$

That is, a person located at -3.0 has a very high probability of incorrectly answering the hardest item on the instrument. The probabilities of incorrect responses to items 3 and 4 would be obtained in a similar fashion. These probabilities are $p(x_3 = 0) = 0.9399$ and $p(x_4 = 0) = 0.9644$.

So far we have the individual item probabilities for a $\theta = -3.0$. To obtain the likelihood of the observed response pattern 11000 requires multiplying the individual item probabilities (step 2). For individuals located at $\theta = -3.0$, the likelihood of observing the response pattern 11000 is given by

$$\begin{aligned} p(x_1 = 1) * p(x_2 = 1) * p(x_3 = 0) * p(x_4 = 0) * p(x_5 = 0) = \\ 0.2497 * 0.0832 * 0.9399 * 0.9644 * 0.9692 = 0.0182 \end{aligned}$$

Stated in words, the likelihood of individuals located at -3.0 providing the response pattern 11000 is about 0.02.

These two steps, calculating first the individual item probabilities and then the joint probability of the pattern, would be repeated for the θ s in the range -3.0 to 3.0 (step 3). Conceptually, the resulting series of probabilities from -3.0 to 3.0 collectively form the *likelihood function* (L). In step 4, the L is examined to determine the *location* of the maximum of the likelihood function. This location is the estimate of the person location ($\hat{\theta}$) that would most likely produce the response pattern 11000 on this mathematics examination, using our model and item parameters.

We may symbolically represent the above steps for calculating a likelihood. Letting \underline{x} represent a response pattern (e.g., $\underline{x} = 11000$), then the likelihood of person i 's response vector, \underline{x}_i , is

$$L(\underline{x}_i | \theta, \alpha, \underline{\delta}) = \prod_{j=1}^L p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})} \quad (2.9)$$

where p_j is short for $p(x_{ij} = 1 | \theta_i, \alpha, \delta_j)$, x_{ij} is person i 's response to item j , $\underline{\delta}$ is a vector containing item location parameters, L is the number of items on the instrument (i.e.,

its length), and “ Π ” is the product symbol. From Equation 2.9 one sees that as the number of items increases, the product of these probabilities will potentially become so small that it will become difficult to represent on any electronic calculation device. Therefore, rather than working directly with the probability, the natural logarithmic transformation of the probability (i.e., $\log_e(p_j)$ or $\ln(p_j)$) is typically used. This transformation results in a summation rather than a multiplication. The use of logs results in a likelihood that is called the *log likelihood function*, $\ln L(\underline{x}_i)$, where $\ln L(\underline{x}_i)$ is

$$\ln L(\underline{x}_i | \theta, \alpha, \delta) = \sum_{j=1}^L [x_{ij} \ln(p_j) + (1 - x_{ij}) \ln(1 - p_j)]. \quad (2.10)$$

A graphical representation of the log likelihood for the pattern 11000 is presented in Figure 2.6. The vertical line in the body of the graph shows that the location of the maximum of the log likelihood occurs at approximately -0.85 (i.e., this is the value that is most likely to result in the response pattern 11000 on this instrument). This value would be the estimated person location for this response pattern (i.e., $\hat{\theta} = -0.85$).

What would the $\ln L$ s look like for the other response patterns that have the same observed score of 2 (e.g., 10100, 01100)? All of these $\ln L$ s exhibit the same form, as seen with the pattern 11000 and with their maxima located at the same θ , but each likelihood is less in magnitude throughout the entire θ continuum than that shown for 11000. Figure 2.7 contains these $\ln L$ s for all 10 patterns with an $X = 2$. This pattern of $\ln L$ s is intuitively appealing because incorrectly answering the easiest three items and correctly answering the hardest two items (i.e., 00011) is not as likely to occur as the reverse pattern 11000. In short, none of the other patterns for $X = 2$ is as likely to occur

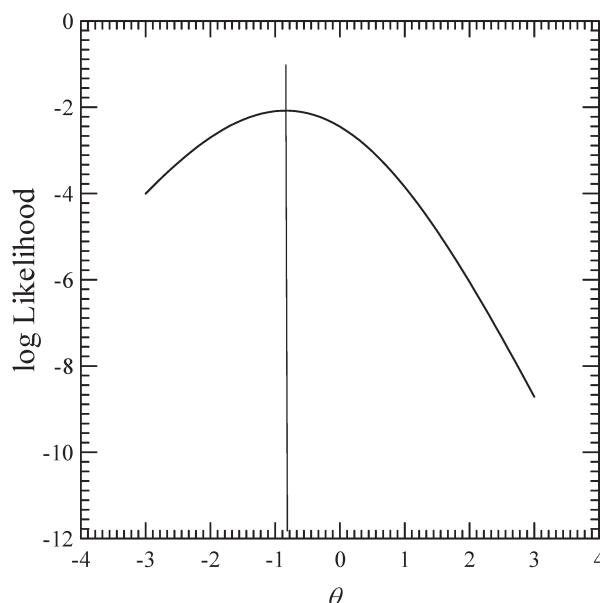


FIGURE 2.6. Log likelihood function for the pattern 11000.

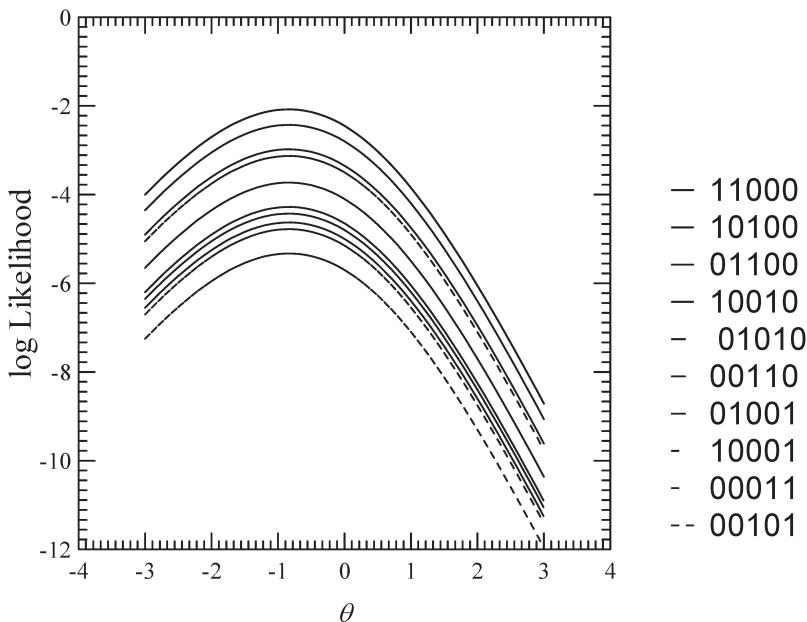


FIGURE 2.7. Log likelihood functions for the all patterns that result in an $X = 2$.

as 11000. Moreover, although different patterns of responses may produce the same X and have varying degrees of likelihood, for the 1PL model a given X yields the same $\hat{\theta}$ regardless of the pattern of responses that produced X . Stated another way, Figure 2.7 shows that for the 1PL model the person's observed score (i.e., $X_i = \alpha \sum x_{ij}$) contains all the information necessary for estimating θ_i or, statistically speaking, X_i is a sufficient statistic for estimating θ_i (cf. Rasch, 1980).¹⁰ In effect, the data shown in Table 2.1 may be collapsed into six observed scores of 0, 1, 2, 3, 4, and 5. As such, and in the context of the 1PL model, the actual pattern of responses that make up each observed score is ignored in our estimate of θ_i when we use the maximum likelihood approach.¹¹ (The actual pattern of responses that make up an X is also ignored in CTT.) We can proceed to obtain the $\hat{\theta}$ s for the remaining patterns in Table 2.1 by determining their lnLs in a fashion similar to that used with 11000.

Although we located the maximum of the lnL by visual inspection, alternative, more sophisticated approaches can be used to find the maximum's location. One of these approaches is Newton's method for *maximum likelihood estimation* (MLE); this approach is discussed in Appendix A. Assume that we use Newton's method to find the MLE $\hat{\theta}$ s for the remaining observed scores. For the individuals who obtained only one correct answer ($X = 1$), their $\hat{\theta} = -1.9876$. That is, these 3099 individuals (i.e., 2280 + 242 + . . . + 184) are located approximately two logits below the zero point on the mathematics proficiency continuum metric. For the observed scores of 2, 3, and 4, the calculated $\hat{\theta}$ s are -0.8378, 0.1008, and 1.1796, respectively. Comparing the MLE $\hat{\theta}$ for $X = 2$ with our visual inspection estimate ($\hat{\theta} = -0.85$; $X = 2$) shows close agreement. We also see that as an individual answers more questions correctly, their corresponding MLE $\hat{\theta}$ increases to

indicate greater mathematics proficiency. However, unlike the observed scores, our $\hat{\theta}$ s are invariant of this particular mathematics examination.

Some Pragmatic Characteristics of Maximum Likelihood Estimates

It may have been noticed that $\hat{\theta}$ s were not provided for people who obtained either a zero score ($X = 0$) or a perfect score ($X = 5$) on the examination. This is because the corresponding log likelihoods do not have a maximum. For instance, the log likelihood for the perfect score ($X = 5$) is presented in Figure 2.8. It can be seen that this log likelihood conforms to what one would expect. There are an infinite number of θ s above 4 that could produce the observed score of 5. Because there is no way of distinguishing which θ among them is most likely, given our data, we do not have a finite $\hat{\theta}$ for a perfect score. In this case, our log likelihood is asymptotic with 0.0, and the estimate of θ would be ∞ . For a zero score, the log likelihood would be the mirror image of the one shown in Figure 2.8 with $\hat{\theta}$ equal to $-\infty$.

When zero and perfect scores are encountered in practice, the various computer estimation programs have different kludges (i.e., “work arounds”) for handling these scores. For example, in the next chapter the perfect and zero scores are modified so that they are no longer perfect and zero, respectively. Alternatively, an estimation approach

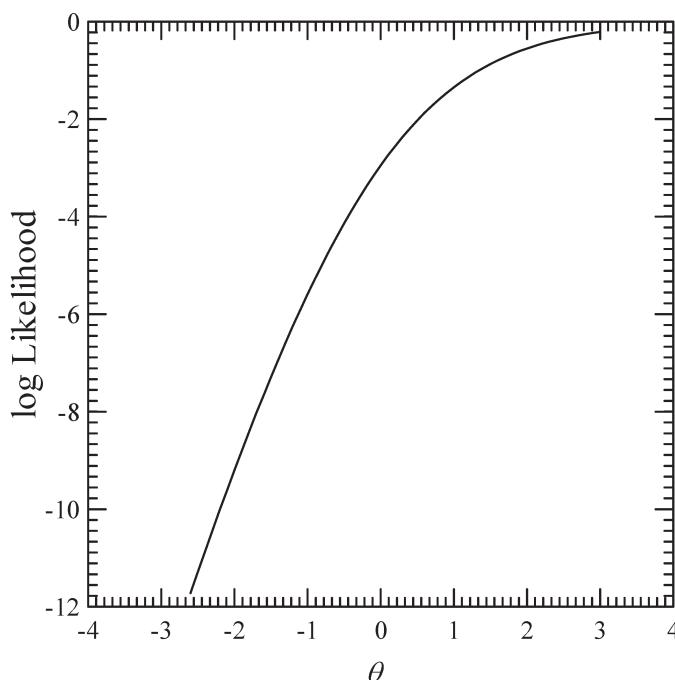


FIGURE 2.8. Log likelihood function for the perfect score of $X = 5$ ($\mathbf{X} = 11111$).

that incorporates ancillary population information can be used to provide $\hat{\theta}$ s for zero and perfect scores. This approach falls within Bayesian estimation and is discussed in Chapter 4.

The Standard Error of Estimate and Information

In general, we can obtain a measure of the amount of error in a statistic's estimate of a parameter. This measure, known as a standard error, is an index of the variability (i.e., the standard deviation) of an estimator with respect to the parameter it is estimating. The larger the value of a standard error, the greater the error and the less certain we are about the parameter's value. Similarly, in IRT our uncertainty about a person's location can be quantified through the estimate's *standard error of estimate* (SEE); we use $\sigma_e(\hat{\theta})$ as shorthand for $\sigma_e(\hat{\theta}|\theta)$.¹² The SEE specifies the accuracy of $\hat{\theta}$ with respect to the person location parameter, θ . When there is a small degree of uncertainty about a person's location, then its SEE is comparatively smaller than when there is a greater degree of uncertainty. The SEE should not be confused with the standard error of measurement used in classical test theory (CTT).¹³

The asymptotic variance error of estimate for $\hat{\theta}$ is

$$\sigma_e^2(\hat{\theta}|\theta) = \frac{1}{\mathcal{E}\left[\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^2\right]} = \frac{1}{\sum_{j=1}^L \frac{(p'_j)^2}{p_j(1-p_j)}}, \quad (2.11)$$

where p_j is given by the IRT model, p'_j is the model's first derivative, and \mathcal{E} is the symbol for expectation (Lord, 1980). Given that the first derivative of the 1PL model is $p'_j = \alpha[p_j(1-p_j)]$, then Equation 2.11 simplifies to

$$\sigma_e(\hat{\theta}|\theta) = \sqrt{\frac{1}{\sum_{j=1}^L \alpha^2 [p_j(1-p_j)]}} \quad (2.12)$$

In practice, $\hat{\theta}$ is substituted for θ in the IRT model. For example, for the person location estimate of -0.8378 , our SEE is 0.9900 (i.e., roughly a full logit).

Table 2.2 contains the MLE $\hat{\theta}$ s and their corresponding SEEs. The magnitude of the SEE is influenced not only by the quality of the items on the instrument, but also by the instrument's length. The addition of items similar to those on the instrument will lead to a decrease in the standard errors of $\hat{\theta}$. For example, if we lengthen our example mathematics test to 20 items by quadrupling the five-item set (i.e., four items at -1.9 , four items at -0.6 , etc.), then our SEE for $\hat{\theta} = -0.8378$ decreases to 0.4950 .

We can use our SEEs to create a maximum likelihood confidence limit estimator (Birnbaum, 1968) by

$$(1-\alpha)\%CB = \hat{\theta} \pm z_{(1-\alpha/2)} \sigma_e(\hat{\theta}) \quad (2.13)$$

Equation 2.13 tells us the range within which we would expect θ to lie $(1-\alpha)\%$ of

TABLE 2.2. MLE $\hat{\theta}$ s and Their Corresponding SEEs for the Different Xs

vv	$\hat{\theta}$	$s_e(\hat{\theta})$	95% CB	Number of Iterations
0	$-\infty$.	.	∞
1	-1.9876	1.2002	-4.3401, 0.3648	4
2	-0.8378	0.9900	-2.7783, 1.1026	3
3	0.1008	0.9717	-1.8037, 2.0053	3
4	1.1796	1.1562	-1.0864, 3.4457	3
5	∞	.	.	∞

Note. CB, confidence band.

the time.¹⁴ For the example's observed score of 2 with $\hat{\theta} = -0.8378$, the 95% confidence band would be $[-0.8378 \pm 1.96 \cdot 0.9900] = [-2.7783, 1.1026]$. That is, we would expect the θ that produced an $X = 2$ on the instrument to lie within this interval 95% of the time. The width of the confidence band is directly related to the degree of uncertainty about a person's location. A narrow interval indicates comparatively less uncertainty about a person's location than a wider interval.

So far we have viewed the estimation of a person's location from the perspective of how uncertain we are about the person's location. We can also take the complementary perspective. That is, how certain are we about a person's location, or, similarly, how much *information* do we have about a person's location? From this perspective, the confidence band's width is inversely related to the information we have for estimating a person's location with an instrument. A narrow interval indicates comparatively more information for estimating a person's location than does a wider interval. If we take the reciprocal of Equation 2.11, we obtain an expression that directly specifies the "amount of information to be expected in respect of any unknown parameters, from a given number of observations of independent objects or events, the frequencies of which depend on that parameter" (Fisher, 1935, p. 18).

Because the instrument's items are our "observations" and given the conditional independence assumption, Fisher's idea about information may be applied to quantify the amount of information that items as well as the instrument provide for estimating the person location parameters. Following Fisher's use of " I " to represent the concept information, then an estimator's information equals the reciprocal of Equation 2.11

$$I(\theta) = -\mathcal{E}\left[\frac{\partial^2}{\partial \theta^2} \ln L\right] = \frac{1}{\sigma_e^2(\theta)}. \quad (2.14)$$

By substitution of Equation 2.11 into Equation 2.14, we obtain the *total information* ($I(\theta)$) provided by the instrument for estimating θ

$$I(\theta) = \frac{1}{\sigma_e^2(\theta)} = \sum_{j=1}^L \frac{(p'_j)^2}{p_j(1-p_j)}, \quad (2.15)$$

where all terms have been defined above. Equation 2.15 is also referred to as *test information* or *total test information*. Unlike the concept of reliability that depends on both instrument and sample characteristics, an instrument's total information is a property of the instrument itself (cf. Samejima, 1990). In this book, the term *total information* is used in lieu of *test information* or *total test information* to reflect the fact that the instrument may not necessarily be a test.

Equation 2.15 specifies how much information an instrument provides for separating two distinct θ s, θ_1 and θ_2 , that are in proximity to one another. By analogy, in simple linear regression the steeper the slope of the regression line, the greater the difference between the predicted values for two different predictor values than when the slope is less steep. For example, imagine the slope is 0 in one case and 0.9 in another case. In the former situation one would predict the same value for two different predictor values, whereas in the latter one would predict two different values for two different predictor values. The numerator of Equation 2.15 is the (squared) slope, whereas the denominator is a reflection of the variability at the point at which the slope is calculated. Therefore, less variability (i.e., greater certainty) at the point at which one calculates the slope combined with a steep slope provides more information for distinguishing between θ_1 and θ_2 than if one had more variability and/or a less steep slope. Moreover, Equation 2.15 shows that, all things being equal, lengthening an instrument leads to a concomitant increase in precision for estimating person locations.

An instrument's total information reflects that each of the items potentially contributes some information to reduce the uncertainty about a person's location *independent* of the other items on the instrument. It is because of this independence that we can sum the individual item contributions to obtain the total information (Equation 2.15). This individual item contribution is known as *item information*, $I_j(\theta)$ (Birnbaum, 1968)

$$I_j(\theta) = \frac{(p'_j)^2}{p_j(1-p_j)} \quad (2.16)$$

(The subscript j on I signifies item information, whereas the lack of a subscript indicates total information.) Therefore, total information equals the sum of the item information, $I(\theta) = \sum I_j(\theta)$.

For the 1PL model, Equation 2.16 simplifies to

$$I_j(\theta) = \alpha^2 p_j(1-p_j). \quad (2.17)$$

Because the product $p_j(1-p_j)$ reaches its maximum value when $p_j = (1-p_j)$, the maximum item information for the 1PL model is $0.25\alpha^2$. Figure 2.9 shows an example item information function for an item located at 0.35 based on the Rasch model (i.e., $\alpha = 1.0$). As can be seen, the item information function is unimodal and symmetric about the item's location with a maximum value of 0.25.

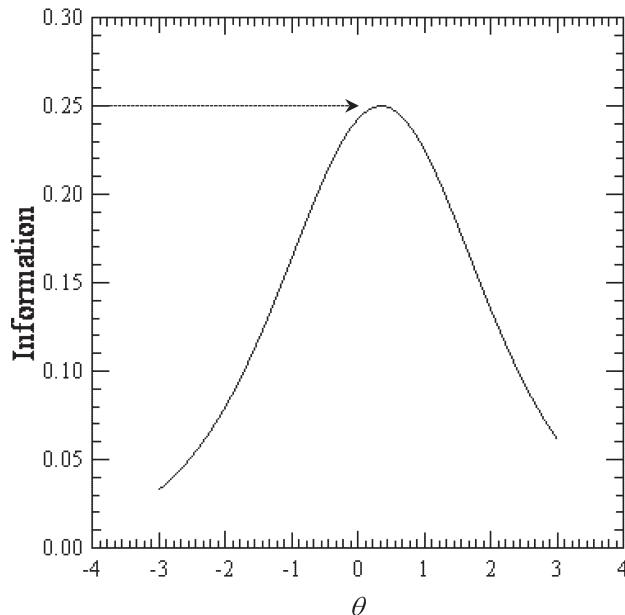


FIGURE 2.9. Item Information for item ($\delta = 0.35$, $\alpha = 1.0$).

With the 1PL model all the items exhibit the same pattern seen in Figure 2.9. Namely, (1) an item provides its maximum information at its location, (2) the item information function is unimodal and symmetric about δ , and (3) all items on an instrument provide the same maximum amount of information of $\alpha^2 0.25$ at their respective locations. We now apply these concepts to our instrument.

An Instrument's Estimation Capacity

The likelihood principles outlined above for person estimation can also be applied to the estimation of item locations. The MLE of item locations is presented in Appendix B. Let us say that, using MLE, we estimated the five item locations. Our estimates, $\hat{\delta}$ s, are presented in Table 2.3 along with their corresponding standard errors. As we see, item 1 is located at ($\hat{\delta}_1 = -1.90$), item 2 is located at $\hat{\delta}_2 = -0.60$, and so on; $\alpha = 1$. In general, item 1 would be considered to be an easy item. The most difficult item on the mathematics instrument is item 5 with an estimated location of 0.45. (Because we are using estimated person parameters in calculating the standard error, our estimate of $\sigma_e(\hat{\delta})$ is given by $s_e(\hat{\delta})$.)

Figure 2.10 contains the items' corresponding IRFs. Because the items have a common α and α is related to an IRF's slope, it is not surprising the IRFs are parallel to one another. One also sees that the probability of a correct response is 0.5 at the items' estimated locations (i.e., an item's location corresponds to the IRF's point of inflection).

To obtain an idea of how well an item and the entire instrument can estimate person

TABLE 2.3. MLE $\hat{\delta}$ s and Their Corresponding SEEs for the Five-Item Instrument

Item	$\hat{\delta}$	$s_e(\hat{\delta})$	Number of iterations
1	-1.90	0.0246	4
2	-0.60	0.0183	4
3	-0.25	0.0183	4
4	0.30	0.0198	4
5	0.45	0.0209	4

locations, we examine the item and total information. Figure 2.11 shows the information provided by each item ($I_j(\theta)$) and by the instrument ($I(\theta)$) as a function of θ . As can be seen, item 1 (nonbold solid line) provides its maximum information for estimating θ at its location of -1.90. As we move away from this point, in either direction, the item provides progressively less information about θ . In fact, above 2.0 this item provides virtually no information for estimating persons located at and above 2.0. Therefore, using this item and others like it to estimate individuals with $\theta > 2.0$ would not yield precise estimates, and the corresponding standard errors would be comparatively large.

The total information function (labeled “Total”) shows the instrument provides its maximum information for estimating θ in a neighborhood around 0.70. As we progress

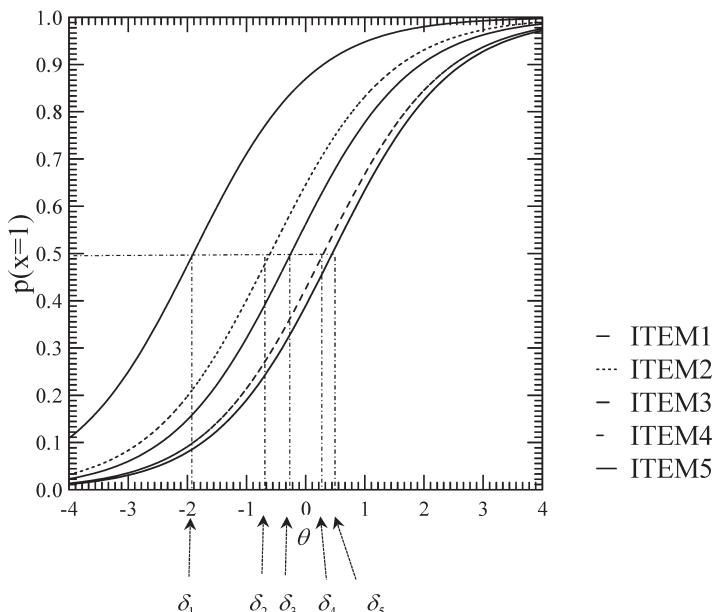


FIGURE 2.10. IRFs for all five items on the Mathematics Instrument.

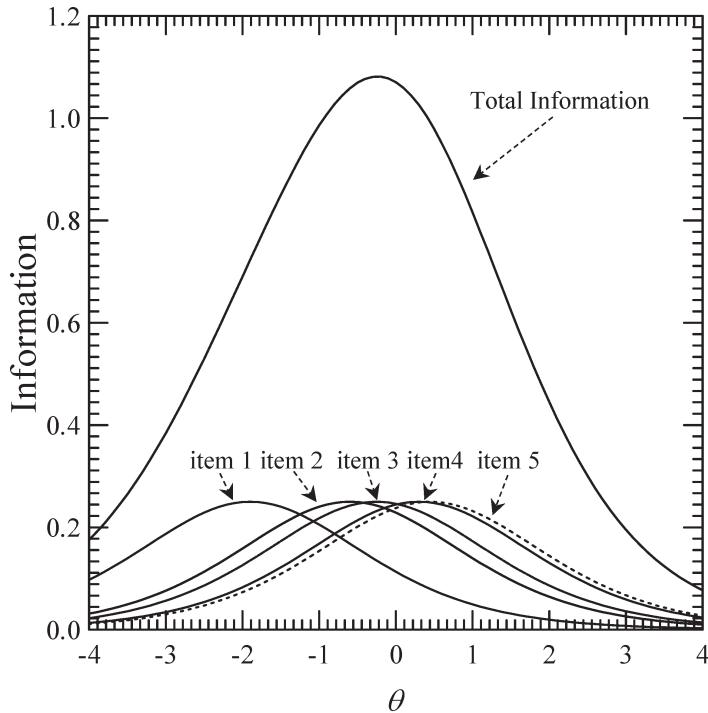


FIGURE 2.11. Item and total information functions for the Mathematics Instrument.

away from this neighborhood, the instrument provides less information for estimating θ so that, for example, at $\theta = 3.0$ the instrument provides one half less information for estimating θ than at 0.70.¹⁵ Recalling the inverse relationship between information and SEE, we find that this observed decrease in information is commensurate with an increase in the standard errors for the $\hat{\theta}$ s located away from 0.70.

This knowledge of how an instrument will behave in estimating person locations permits the design of an instrument with specific estimation properties. For example, if we are interested in providing better estimation below 0.7, then we need to improve the operational range of the test. We can do this by the addition of one or more items at the lower end of the continuum to increase the information about individuals located at the lower end of the continuum. If it were necessary to restrict our instrument to five items, then we might remove one or more items from the instrument that provide redundant information. For example, we might consider removing item 4 ($\delta_4 = 0.30$) because it and item 5 ($\delta_5 = 0.45$) are providing somewhat redundant information about examinee location. In contrast, if we desired to provide better estimation above 0.7, then items located around 0.7 and greater would be included in the instrument. In this fashion, we can design an instrument to measure along a wide range of the continuum or, conversely, very precisely in a narrow range by adding items located within the range of interest.

To facilitate designing an instrument with specific estimation properties, we can

specify a *target total information function* for the instrument. This function can provide a blueprint for the area(s) on the continuum for which we wish to provide a high degree of precision for estimating person locations. For instance, consider a certification or licensure situation that involves a (decision) cutpoint above which a person would be, for example, certified and below which the person would not be certified. In this case, it would be desirable to have an instrument with a total information function that is peaked at the cutpoint. This instrument would have enhanced capacity for distinguishing between individuals who were located near the cutpoint. To achieve this goal, we would add items whose information maxima were at or near the cutpoint. Moreover, this greater information would reduce our SEEs at the cutpoint and thereby reduce the confidence band width (Equation 2.13) used to decide whether someone is above or below the cutpoint. (That is, we could use the interval estimate from Equation 2.13, not the point estimate $\hat{\theta}$, for deciding whether an individual is above or below the cutpoint.)

Alternatively, we may wish to have equiprecise estimation across a range of θ (i.e., “constant” SEE). In this case, the corresponding target total information function would resemble a rectangular (i.e., uniform) distribution across the θ range. To achieve this objective, we would add items whose information maxima are located at different points along the θ range of interest.

In summary, information functions may be used to design an instrument with specific characteristics. This capability takes advantage of the fact that items and persons are located on the same continuum, as well as our capacity to assess the amount of information for estimating person locations based solely on the item parameter estimates. Success in developing an instrument whose observed total information function is similar to the target information function depends on having an adequate pool of items to work with and imposing constraints on the item selection to ensure the resulting instrument has validity with respect to the construct of interest. Theunissen (1985), van der Linden and Boekkooi-Timmeringa (1989), and van der Linden (1998) contain detailed information for automating instrument development using a target instrument information function and taking into consideration content representation; also see Baker, Cohen, and Baarmish (1988) as well as Luecht (1998). In general, it is not possible to use CTT to design an instrument in this fashion. Of course, after developing the instrument, it would still be necessary to perform a validation study for the instrument.

Summary

Although for some individuals the Rasch and 1PL models reflect different perspectives for performing measurement, the models are mathematically equivalent. For some, the Rasch model is the standard by which to create a measurement device. From this perspective, for the data to be useful for measurement they must follow the model. Data that do not exhibit fit with the model are seen as suspect and may need to be discarded. In contrast, the 1PL model may be viewed as representing a statistical approach of trying to model response data. In this case, if one has model–data misfit, it is the model that

is seen as suspect. For both models and for IRT in general, the graphical representation of the predicted probabilities of a response of 1 on an item are referred to as an item response function.

Because the 1PL model states that items have a common discrimination parameter (α), the corresponding IRFs are parallel to one another. The item's location (δ) on the latent continuum is defined by the location of the IRF's point of inflection.¹⁶ In addition, we assume the construct of interest is unidimensional, the data are consistent with the model's function form, and the responses are conditionally independent. The unidimensionality assumption states that the observations may be explained solely on the basis of a single latent person trait, θ , whereas the functional form assumption says the response data for an item follow an ogival pattern. In the context of the 1PL model, the conditional independence assumption states that the responses to an item are independent of the responses to another item, conditional on person location. The tenability of these assumptions for a data set needs to be examined.

In contrast to CTT, in IRT both persons and items are located on the same continuum. Although items are allowed to vary in their locations, an item's capacity to differentiate among individuals is held constant across items. This capacity is captured by the item discrimination parameter. For the Rasch model $\alpha = 1.0$ and for the 1PL model α may be equal to some constant other than 1.0.

With the 1PL model, the sum of person i 's responses (observed score) is a sufficient statistic for estimating their location, and the sum of the responses to an item j (item score) is a sufficient statistic for estimating the item's location. All individuals who receive the same observed score will obtain the same estimated person location ($\hat{\theta}$), and all items that have the same item score will receive the same estimated location ($\hat{\delta}$). The accuracy of the $\hat{\theta}$ and $\hat{\delta}$ is indexed by their respective standard errors. Smaller standard errors of estimate reflect greater accuracy than do larger standard errors. Moreover, one's uncertainty of each parameter is reflected in the concept of information. In general, each item provides information for estimating a person's location. The sum of the item information functions is the instrument's total information function. The concept of total information can be used to design instruments with specific psychometric properties.

In this chapter, we conceptually presented maximum likelihood estimation; Appendices A and B contain a more formal treatment for estimating the person and item parameters, respectively. In the next chapter, the estimation of person and item parameters is further developed, using the likelihood principle presented above. Furthermore, we demonstrate a process that can be used in practice for obtaining IRT parameter estimates. As part of this process, we assess some of IRT's assumptions and examine model–data fit.

Notes

1. There are other ways of conceptualizing the relationship between the person and item locations that do not involve taking their difference. For instance, Ramsay (1989) used the ratio of the person and item locations (i.e., θ/δ). Although Ramsay's

model is slightly more complicated than the standard approach of examining the difference between θ and δ , it does have the benefits of allowing different estimates of θ for different response patterns that yield the same observed score, as well as a way of accounting for examinee guessing.

2. Equation 2.2 is a special case of the general unidimensional Rasch model (Wright, 1980). The general Rasch model specifies the probability of a response, x , of falling into one of item j 's m_j response categories

$$p(x_{ij} = x | \theta_i, \delta_j, \underline{\kappa}, \underline{x}) = \frac{e^{[x(\theta_i - \delta_j) - \kappa_x]}}{\sum_{v=0}^{m_j} e^{[\underline{x}_v(\theta_i - \delta_j) - \kappa_v]}}, \quad (2.18)$$

where $\underline{\kappa}$ is a vector of category coefficients with its v th element identified as κ_v , \underline{x} is a vector of category values with its v th element identified as \underline{x}_v , and all other terms are as defined in Equation 2.2. $\underline{\kappa}$ and \underline{x} are of dimension $m_j + 1$ with x taking on values $(0, 1, \dots, m_j)$. This model is applicable to dichotomous ($m_j = 2$) and polytomous ($m_j > 2$) data. (This model should not be confused with Zwinderman's (1991) generalized Rasch model.) The general Rasch model subsumes the partial credit model (Masters, 1982) and the rating scale models (Andrich, 1978b, 1978c), as well as the Rasch Poisson Counts model [see Wright (1980) and Masters (1988)]. The Rasch Poisson Counts model (Rasch, 1980) is focused on tests of multiple attempts on an item (Jansen, 1994). The partial credit model and the rating scale models are discussed in Chapter 7.

3. Some treatments of the Rasch model (e.g., Rasch, 1961; Wright, 1968) use the concept of an item's "easiness" (E_j) to represent an item's location. In these cases, easiness may be transformed to item difficulty, δ , by $\delta_j = \ln(E_j)$ (i.e., $E_j = e^{-\delta_j}$).
4. To obtain this $\hat{\delta}_j$, we use an approximation strategy based on the item's *traditional item difficulty*, P_j ; this approach is discussed in Chapter 5 and Appendix C. Specifically, δ_j is equal to the z -score that delimits an area above it that equals P_j . The P_j for the item in Figure 2.2 is 0.644. From, for example, a standard unit normal curve table, we find that a $z = -0.37$ corresponds to the point above which lies 0.644 of the area under the normal curve. Therefore, the estimate of the item's location is -0.37 . A more sophisticated approach for estimating an item's location is presented in Appendix B.
5. The first derivative, p'_j , of Equation 2.7 is

$$p'_j = \alpha[p_j(1-p_j)] = \alpha \left[\frac{e^{\alpha(\theta-\delta)}}{[1+e^{\alpha(\theta-\delta)}]^2} \right] \quad (2.19)$$

and because by definition α is defined at $\theta = \delta_j$, p'_j simplifies to

$$p'_j = \alpha \left[\frac{e^{\alpha(\theta-\delta)}}{[1+e^{\alpha(\theta-\delta)}]^2} \right] = \alpha \left[\frac{e^0}{[1+e^0]^2} \right] = 0.25\alpha. \quad (2.20)$$

Therefore, strictly speaking, α in Equation 2.7 is proportional to the slope of the tangent line to the IRF at δ_j . Note that, by convention, we define the IRF's slope to be the one calculated at the item's location (i.e., an IRF has different slopes at different points along the function).

6. The IRFs are parallel when α is constant across items with different locations *and* when the lower asymptote of the corresponding IRFs is constant (e.g., the IRF lower asymptotes equal 0.0). Moreover, it is the tangent lines to the IRF at δ_j for the items that are parallel.
7. This philosophy is similar to the Guttman Scalogram (Guttman, 1950) technique for measuring, for example, attitudes. In scalogram analysis, if one can successfully apply the technique, then the resulting scale is unidimensional and has certain known properties. However, there is no guarantee that one will be able to successfully apply scalogram analysis to a particular data set. In short, simply because one would like to measure a particular construct does not necessarily mean one will be successful. This standard is greater than simply asking a series of questions: The data must fit a scalogram in order for the creation of an instrument to be considered successful. (Strictly speaking, there is some latitude.) Similarly, using the Rasch model as the standard for being able to measure a construct means the data must fit the Rasch model. Whether the unidimensional scale produced by the Rasch model is meaningful or useful is a validity question. The Rasch model differs from the Guttman Scalogram model most notably in that the Rasch model is a probabilistic model, whereas the scalogram model is deterministic.
8. McDonald (1979, 1981; McDonald & Mok, 1995) has asserted that there are two principles of conditional independence. The first is the *strong principle of conditional (local) independence*, and the second is the *weak principle of conditional (local) independence*. The strong principle is the one defined by Lord and Novick (1968); that is, after taking into account (conditioning on) all the relevant latent variables, the item responses for *all* subsets of items are mutually statistically independent. The weak principle states that the covariability between *two* item vectors is zero after conditioning on all the relevant latent variables (i.e., pairwise conditional independence). McDonald refers to these two forms as *principles* rather than assumptions because “the (strong or weak) principle of local independence is not an assumption, but serves to provide the mathematical definition of latent traits” (McDonald & Mok, 1995, p. 25). The existence of weak conditional independence is a necessary, but not sufficient, condition for the existence of strong conditional independence.

The weak principle of conditional independence is the basic assumption that underlies common factor analysis (McDonald, 1979). This connection with factor analysis and the capability of factor analysis to distinguish between major and minor factors provides another way of looking at the dimensionality assumption. *Essential dimensionality* (d_E ; Stout, 1990) is the minimal number of major factors necessary for a weakly monotone latent model to achieve *essential independence* (EI). Stout (1990) states that EI exists when the conditional covariances between items, on average, approach zero as test length becomes infinite. When d_E equals 1, then

one has *essentially unidimensionality*. Stout (1987) developed a statistical test, T , to detect departure from (essential) unidimensionality in a data set. This approach is implemented in DIMTEST; DIMTEST is available from the Psychometric Software Exchange (<http://www.psychsoft.soe.vt.edu>). Nandakumar (1991) and Hattie, Krokowski, Rogers, and Swaminathan (1996) contain readable presentations of the DIMTEST approach to assessing dimensionality.

9. This idea of logically consistent response patterns or *ideal response patterns* is seen in a perfect Guttman scale (Guttman, 1950) or in Walker's (1931) *unig test*. A unig test consists of observed scores, X , composed of the correct answers to the L easiest items. Moreover, the observed score $X + 1$ contains the correct answers of score X plus one more. In the context of the 1PL model, the condition for an ideal response pattern for person i is $x_{ij} \geq x_{iv}$ when $p_{ij} \geq p_{iv}$, where j and v are two items. For example, assuming three items are ordered from easy to hard, then the logically consistent patterns are 000 (all items incorrect), 100 (easiest item correct, harder items incorrect), 110 (hardest item incorrect, easier items correct), and 111 (all items correct). These ideal response patterns are also known as *Guttman patterns* or *conformal patterns*. Although our instrument does not conform to a unig test (or a perfect Guttman scale), the ideal response patterns (i.e., 00000, 10000, 11000, 11100, 11110, 11111) have the largest frequency for a given observed score, X . For instance, for $X = 1$ (i.e., the patterns 10000, 01000, 00100, 00010, 00001), the 10000 pattern has the largest frequency. There are

$$C_x^L = \frac{L!}{X!(L-X)!}$$

different pattern combinations for a given X score.

10. According to Fisher (1971b), "if θ [is] the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ , which the sample supplies, and θ_2 any other statistic, then . . . when θ_1 is known, knowledge of the value of θ_2 throws no further light upon the value of θ " (pp. 316–317). By definition, a statistic or estimator (y) for a parameter is *sufficient* if it contains all of the information regarding the parameter that can be obtained from the sample (i.e., no further information about the parameter can be obtained from any other sample statistic (e.g., y') that is functionally independent of y). In short, a sufficient statistic represents a form of data reduction that preserves the information contained in the data. To demonstrate the existence of sufficient statistics for the Rasch model, we follow Wright and Stone (1979). Starting with an $N \times L$ data matrix \underline{X} whose entries x_{ij} represent the binary response on the j th item by the i th person, then the marginal totals (i.e., row and column sums) for \underline{X} are person i 's observed score (i.e., row i in \underline{X})

$$X_i = \sum_{j=1}^L w_j x_{ij}$$

and item j 's item score (i.e., column j in \underline{X})

$$q_j = \sum_{i=1}^N v_i x_{ij}$$

where $w_j = 1.0$ and $v_i = 1.0$.

The probability of the entire data matrix \underline{X} is given by taking the products across the N persons and the L items

$$p_j = \prod_{i=1}^N \prod_{j=1}^L \frac{e^{x_{ij}(\theta_i - \delta_j)}}{1 + e^{(\theta_i - \delta_j)}}$$

Converting these products to sums and then factoring the numerator, we have

$$p_j = \frac{\exp(\sum_i X_i \theta_i) \exp(-\sum_j q_j \delta_j)}{\prod_{i=1}^N \prod_{j=1}^L [1 + \exp(\theta_i - \delta_j)]}$$

where we use “ $\exp[z]$ ” instead of “ e^z ” to simplify presentation. This numerator has the form of sufficient statistics: one for estimating θ_i and the other for estimating δ_j . (See Fisher [1971b] for the condition that must be satisfied for the existence of a sufficient statistic.) Specifically, the observed score X_i ($= \sum_j x_{ij}$) is a sufficient statistic for estimating θ_i , and the item score q_j ($= \sum_i x_{ij}$) is a sufficient statistic for estimating δ_j (also see Rasch, 1980).

The sufficient statistic X_i is the sum of person i 's responses to the items and does not involve item characteristics (e.g., item locations or discriminations). This is why the 10 patterns with an $X_i = 2$ (Figure 2.7) all have the same $\hat{\theta}$. Similarly, the sufficient statistic q_j is the sum of the responses on item j and involves only item responses and not person characteristics.

11. This does not mean that an alternative estimation procedure that does not involve the criterion of maximizing the likelihood function will also yield the same $\hat{\theta}$ for all response patterns that produce a given observed score, X . For example, the minimum χ^2 estimation method seeks to estimate the parameters that minimize a χ^2 criterion (i.e., maximize data–model fit). (See Linacre [2004] or Baker and Kim [2004] for more information on this estimation method.) Utilizing this procedure, “persons with different response patterns, but with the same raw <observe> score, obtain different measures $\langle\hat{\theta}\rangle$ s characterized by different standard errors, and similarly for items” (Linacre, 2004, p. 6).
12. Strictly speaking, one only has the sample standard error of estimate ($s_e(\hat{\theta})$); that is, an estimate of $\sigma_e(\hat{\theta})$.
13. The standard error of measurement is a *global* measure of error across the manifest variable score metric for a given sample. It may (and most likely will) over- or underestimate the degree of error at different points along the score metric (Feldt & Brennan, 1989; Livingston, 1982). In our example, there are six possible X_i s, but there is only one standard error of measurement. One could calculate a *conditional standard error of measurement* at each X_i between the zero and perfect scores that would provide a more accurate reflection of the error at that particular score than

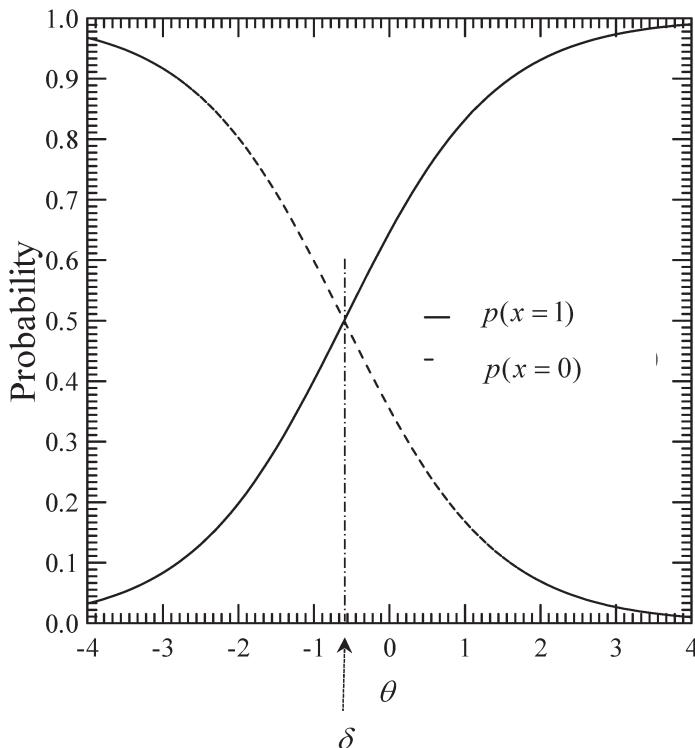


FIGURE 2.12. Probability functions for a response of 1 and a response of 0.

would the standard error of measurement. See Livingston (1982), Kolen, Hanson, and Brennan (1992), and Kolen, Zeng, and Hanson (1996) for approaches to calculating a conditional standard error of measurement. These conditional standard errors are the same for everyone who obtained a given X_i .

14. The use of z in this formula is based on the fact “that a maximum likelihood estimator has approximately (asymptotically) the normal distribution with mean θ , the true ability value, and variance $1/I(\theta)$, under conditions satisfied by most of the models” (Cramér, 1946, p. 500; cited in Birnbaum, 1968, p. 457).
15. As Lord (1980) points out, “information is not a pure number; the units in terms of which information is measured depend on the units used to measure ability” (p. 87). Therefore, the foregoing description of the locations of item and total information maxima as well as the shape of the information function is tied to the particular θ metric used and should not be considered absolute.
16. The item’s location (δ) on the latent continuum is the point of intersection of the IRF (i.e., $p(x_j = 1 | \theta, \delta)$) and the function for the response of 0 (i.e., $p(x_j = 0 | \theta, \delta) = 1 - p(x_j = 1 | \theta, \delta)$). Figure 2.12 shows these two functions. As can be seen, these two functions are mirror images with the sum of their respective probabilities equal to 1 conditional on θ .

3

Joint Maximum Likelihood Parameter Estimation

In Chapter 2 we introduced the Rasch model. As part of its presentation, we discussed using the likelihood of the observed responses for estimating the model's parameters; also see Appendices A and B. Multiple approaches make use of the likelihood function for estimating item and person parameters. In this chapter we present one approach, and in Chapter 4 a second procedure, marginal maximum likelihood estimation, is discussed.

In practice, neither the person nor the item location parameters are known. As such, we conceptually introduce a procedure for the simultaneous estimation of both item and person parameters. We then present an example of applying this estimation procedure and some of the steps involved in assessing model–data fit, such as dimensionality and invariance assessment.

Joint Maximum Likelihood Estimation

Various strategies have been developed to solve the problem of estimating one set of parameters (e.g., the person set) without knowledge of another set of parameters (e.g., the item set). One of these approaches maximizes the joint likelihood function of both persons and items to simultaneously estimate both person and item parameters. This strategy is called *joint maximum likelihood estimation* (JMLE). To arrive at the joint likelihood function for persons and items, we begin with the likelihood function for persons. Assuming conditional independence and (without loss of generalizability) dichotomous responses, the probability of a person's responses is simply the product of the probability of the responses across an instrument's items. Symbolically, for the dichotomous case, this can be represented for an L-item long instrument as

$$p(\underline{x} | \theta, \alpha, \underline{\delta}) = \prod_{j=1}^L p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})} \quad (3.1)$$

The term $p(\underline{x}|\theta, \alpha, \delta)$ is the probability of the response vector, \underline{x} , conditional on the person's location (θ), item discrimination (α), and on a vector of item location parameters, δ (i.e., $\delta = (\delta_1, \dots, \delta_L)$). The probability for item j , p_j , is calculated according to a particular model (e.g., the 1PL model).

To obtain the joint likelihood function, L , across both persons and items, one multiplies Equation 3.1 across the N persons

$$L = \prod_{i=1}^N \prod_{j=1}^L p_j(\theta_i)^{x_{ij}} (1 - p_j(\theta_i))^{(1-x_{ij})} \quad (3.2)$$

Recall from Chapter 2 that to avoid numerical accuracy issues, the likelihood function is typically transformed by using the natural log ($\ln(\bullet)$). Therefore, by applying the natural log transformation to Equation 3.2, we obtain the joint log likelihood function

$$\ln L = \sum_{i=1}^N \sum_{j=1}^L [x_{ij} \ln(p_j(\theta_i)) + (1 - x_{ij}) \ln(1 - p_j(\theta_i))] \quad (3.3)$$

The values of the θ s and δ s that maximize Equation 3.3 are the person and item parameter estimates, respectively. These estimates are determined by setting the first derivatives of $\ln L$ to zero, similar to what is presented in Appendices A and B.

This strategy of maximizing the joint likelihood function proceeds in a series of steps and stages. (For simplicity we assume the Rasch model in describing these steps.) In step 1 the item locations are estimated using provisional estimates of person locations. These provisional person location estimates are treated as "known" for purposes of estimating the item locations. The estimation of item locations is done first because, typically, one has substantially more persons than items, and thereby there is more information for estimating the item locations. Because the estimation of one item's parameter does not depend on the parameters of other items, the items are estimated one at a time. In step 2 these estimates are treated as "known" and are used for estimating person locations. Each person's location is estimated independently of those of the other individuals.

In step 1 the estimation of the item locations used provisional estimates of person locations. However, after step 2 these provisional person estimates have been improved, and these improved estimates should lead to more accurate item location estimates. Therefore, in our second stage, step 1 is repeated using the improved estimates of person locations from step 2 to obtain better estimates of the item locations. With these improved item location estimates, step 2 is repeated using the improved item location estimates to improve the person location estimates. This "ping-pong" process continues until successive improvements in the person and item location estimates are considered "indistinguishable" from one another (i.e., these improvements are less than the *convergence criterion*; see Appendix A). See Baker and Kim (2004) and Lord (1980) for further details and the relevant equations.

On occasion one encounters the term *unconditional maximum likelihood estimation* (UCON). UCON is a synonymous term for JMLE. The term UCON has been used by Ben Wright (e.g., Wright & Stone, 1979) to distinguish this approach from another estimation approach used with the Rasch model called *conditional maximum likelihood*

estimation (CMLE; Andersen, 1972). (CMLE takes advantage of the separability of the person and item parameters in the Rasch model to condition the likelihood function on the Rasch model's sufficient statistics. The result provides consistent maximum likelihood estimators of δ [Andersen, 1972]. CMLE can only be used with the Rasch model and its various extensions.)

JMLE (UCON) is used in estimation programs such as WINSTEPS (Linacre, 2001a, 2005), BIGSTEPS (Linacre & Wright, 2001), FACETS (Linacre, 2001b), and Quest (Adams & Khoo, 1996), as well as in LOGIST (Wingersky, Barton, & Lord, 1982). CMLE is one of two estimation techniques available in OPLM (Verhelst, Glas, & Verstralen, 1995) and is available with the R packages eRm (Mair, Hatzinger, & Maier, 2018) and TAM (Robitzsch, Kiefer, & Wu, 2020). More technical information about JMLE and CMLE estimation may be found in Baker and Kim (2004); Baker and Kim (2017) present the R programming to implement the estimation algorithm as well as various plotting functions (e.g., for IRFs).

Indeterminacy of Parameter Estimates

With CTT, the trait score's metric is determined by the expectation of the observed scores. However, in IRT this is not the case. To understand why this is so, consider our development of the Rasch model. In developing the model, our concern is only with the distance between the person and an item's locations, $(\theta - \delta)$. If $\theta = 2$ and $\delta = 1$, then this distance is 1.0 and the probability of a response of 1 with the Rasch model would equal 0.7311. However, for any θ and δ whose difference equals 1.0 the Rasch model would result in exactly the same probability of a successful response (e.g., if $\theta' = 50$ and $\delta' = 49$, then $p_j = 0.7311$). This raises the question, "Should the person be located at 2 and the item located at 1, or should the person be located at 50 and the item located at 49?" In other words, because multiple values of θ and δ lead to the same probability, the continuum's metric is *not absolute*, but rather *relative* and *nonunique*. In IRT we have an indeterminacy in the origin and unit of measurement of the metric. In short, our metric is unique *only* up to a linear transformation. For our example, the linear transformation for the θ s is $\theta^* = \theta(1) + 48$, and for the δ s it is $\delta^* = \delta(1) + 48$. This property is referred to as the *indeterminacy of scale*, *indeterminacy of metric*, or the *model identification problem*.

This indeterminacy requires that we anchor the metric's origin as well as the unit of measurement to estimate the model's parameters. With the Rasch model the unit of measurement is fixed at 1.0, so our focus is on anchoring the metric. Because people and items are located on the same continuum, we need only be concerned with either the person or the item locations to anchor the metric. One method for fixing the metric, *person centering*, sets the mean of the $\hat{\theta}$ s to 0.0 after each step of person location estimation. A second approach, *item centering*, sets the mean of the $\hat{\delta}$ s to 0.0 after each step of item estimation.¹ Although these two approaches will most likely result in different estimates, the relative nature of the metric means that model–data fit will not be adversely affected by the centering approach used.

To understand the effect of this indeterminacy with the JMLE algorithm, recall

that initially we estimate the item locations given provisional estimates of the person locations. In the next step, the person locations are estimated *relative* to the item location estimates. Subsequently, the item locations are re-estimated *relative* to the new estimates of person locations, and so on across the stages. Because with each step the estimation proceeds relative to the improved estimates from the previous step, the metric's mean begins to drift across the steps. Therefore, with JMLE the metric of either the person or item locations needs to be fixed after each step. For instance, and to exemplify item centering, we would first estimate the item locations, calculate the mean $\hat{\delta}$, and then subtract it from each $\hat{\delta}$ to produce a metric centered at 0.0. We would perform this centering after the relevant estimation step. Centering is an example of the application of a linear transformation.

Of the various commonly available IRT programs, WINSTEPS, BIGSTEPS, and FACETS all use item centering. An older program, LOGIST, uses person centering. These programs all use JMLE. Programs that use marginal maximum likelihood estimation, such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), BILOG 3 (Mislevy & Bock, 1997), flexMIRT (Cai, 2013; Houts & Cai, 2013), IRTPRO (Cai, Thissen, & Du Toit, 2020), mirt (Chalmers, 2012, 2015, 2017, 2019), PARSCALE (Muraki & Bock, 2003), and TAM use a variant of person centering based on the posterior latent variable (person) distribution (Baker, 1990); this approach is discussed in Chapter 4.

How Large a Calibration Sample?

The process of obtaining estimates of person and item parameters is called *calibration*. Wright (1977a) stated “that calibration sample sizes of 500 are more than adequate in practice and that useful information can be obtained from samples as small as 100” (p. 224). However, in this latter situation one has less calibration precision than with larger sample sizes (Wright, 1977a) as well as a loss of power for detecting model–data misfit (Whately, 1977).² That is, the degree of model–data misfit that one is willing to tolerate should be taken into consideration when discussing calibration sample sizes. We feel that another consideration in determining calibration sample size should include the sample size requirement(s) of ancillary technique(s), such as methods to be used for dimensionality assessment (e.g., factor analysis). For example, there are various rules of thumb for factor analysis sample sizes, such as the number of persons should be from 3 to 10 times the number of items (these ratios interact with the magnitude of the factor loadings). The 10:1 ratio is a very common suggestion, whereas the smaller ratios are applicable only when the communalities are large (cf. MacCallum, Widaman, Zhang, & Hong, 1999). As such, our calibration sample size is affected by the requirements for performing, for example, factor analysis.

It cannot be stressed enough that sample size guidelines should not be interpreted as hard-and-fast rules. Specific situations may require more or fewer persons than other situations, given the (mis)match between the instrument’s range of item locations and the sample’s range of person locations, the desired degree of estimation accuracy of both items and persons, and pragmatic issues concerning model–data fit, ancillary technique

sample size requirements, estimation technique, the amount of missing data, and generalizability. For instance, if one is interested in establishing norms for a particular population, then the representativeness of the sample would be paramount. This may require obtaining a large sample to formulate a convincing argument to support the utility of the norms. As such, it may be tempting to simply adopt the philosophy that one should have as large a sample size as one can obtain. However, the increased cost of collecting a large sample may not always be justified. For example, if one is performing a survey, then approximately 1200 randomly sampled individuals may be sufficient, provided that the size of the population is large (Dillman, 2000). Another consideration is that the smaller the sample size, the higher the probability, all other things being equal, that everyone will provide the same response (e.g., a response of 0) to one or more items. As a result, the item location is unestimatable. The same problem may occur with “short” instruments. That is, with short instruments there is an increased chance of a person’s providing a response of, for example, 1 for all items. This individual’s location could not be estimated using MLE. Given the foregoing caveats and considerations, a rough sample size guideline is that a calibration should have a few hundred respondents. This should not be interpreted as a minimum, but rather as a desirable target. Certain applications may require more respondents, whereas in other applications a smaller sample may suffice.

Example: Application of the Rasch Model to the Mathematics Data, JMLE, BIGSTEPS

We use the mathematics data introduced in Chapter 2 (Table 2.1) to demonstrate the application of the Rasch model for measurement using JMLE. In Chapter 4, we reanalyze these data using two different programs, BILOG-MG and mirt, to demonstrate marginal maximum likelihood estimation.

For this example, we first use the free BIGSTEPS software program and then we reanalyze the data using the R package mixRasch (Willse, 2011, 2014). The BIGSTEPS and mixRasch calibration programs implement JMLE for the Rasch model and certain extensions of it. In this chapter, we use some of the BIGSTEPS’s features, and in Chapter 7, we introduce additional features where it is used with polytomous data.

One distinguishing feature of BIGSTEPS is its lack of a graphical user interface (GUI; i.e., menus and dialogs). We recognize that many readers are more comfortable using a program with a GUI. Although an alternative program, WINSTEPS, uses a GUI, our decision to use BIGSTEPS is based on it being free (WINSTEPS is not) and on its output being similar enough to WINSTEPS’ that the reader can easily make the transition to WINSTEPS. BIGSTEPS requires the creation of an input control file. This input file can be used with WINSTEPS. As such, the process of creating the control file also transfers to WINSTEPS. (WINSTEPS’ GUI simply facilitates creating the input file.) There is a free “student version” of WINSTEPS called MINISTEPS. Unfortunately, because of its limitations regarding the number of items and cases (75 cases, 25 items), it cannot be used with our data.

The application of IRT for measuring a variable requires three categories of activities.

The first category consists of calibration-related activities. A second category involves model–data fit assessment, and the third category requires obtaining validity evidence. To varying degrees, all three categories inform one another. We use the term *category* rather than, say, *stages*, to emphasize that the corresponding activities are not necessarily conducted in sequence (i.e., category 1 activities precede category 2 activities, etc.). In the following discussion, the primary emphasis is on the first two categories. These two categories are discussed first and are then followed by an example.

Category 1 involves calibration-related activities. Obviously, the entire process begins with the construction of an instrument. The instrument is pilot tested, a traditional item analysis is conducted, the instrument refined (as necessary), and then administered to the individuals of interest; see Bandalos (2018) for more information on instrument construction and item analysis. After this administration, the data are inspected for anomalous responses (e.g., miscoded responses, multiple responses to an item, etc.) and, if necessary, appropriately corrected. This “cleaning” of the data is followed by some of the category 2 activities, and these, in turn, lead to our calibration (category 1).

Category 2 consists of model–data fit activities. Some of these activities are performed prior to the calibration, whereas others occur after the calibration. Although some model–data fit activities transcend individual programs, some approaches for assessing model–data fit are easier to perform with some calibration programs than with others. For example, the assessment of the data’s dimensionality and the examination of invariance of parameter estimates (discussed below) are examples of activities that transcend individual calibration programs. In contrast, specific fit statistics/indices and graphical assessment of fit tend to be particular to a specific program because individual programs provide different fit information.

The following example begins with an assessment of the tenability of the IRT unidimensionality assumption. In short, and in the context of the Rasch model, the question being asked is, “Do our data conform to the unidimensional Rasch model?” After we perform our dimensionality assessment, we proceed to the data’s calibration. As part of this calibration, various fit statistics are produced at both the model and item/person levels. The model-level fit statistics are examined first, followed by the item-level fit statistics. In some cases, model-level *misfit* may be diagnosed by examining the item-level fit statistics. For this example, the final step in our fit examination involves assessing the invariance of the item parameter estimates. This is followed by a brief discussion of obtaining validity evidence for the instrument (i.e., category 3 activities).

Dimensionality Assessment

In assessing dimensionality, our interest is in the number of *content-oriented* factors. The traditional approach for assessing dimensionality involves the factor analysis of a correlation matrix. Whenever one factor analyzes a correlation matrix derived from binary data, there is a possibility of obtaining artifactual factor(s) that are related to the nonlinearity between the items and the common factors. These “factors of curvilinearity” have sometimes been referred to as “difficulty” factors and are not considered to

be content-oriented factors (Ferguson, 1941; McDonald, 1967; McDonald & Ahlawat, 1974; Thurstone, 1938). To avoid extracting these difficulty factors, McDonald (1967) suggests the use of nonlinear factor analysis. Because our data are dichotomous, we use this nonlinear approach for our dimensionality analysis. This nonlinear strategy is implemented in the program NOHARM (Fraser, 1988; Fraser & McDonald, 2003, 2012). NOHARM performs well in dimensionality recovery studies (e.g., De Champlain & Ges-saroli, 1998; Finch & Habing, 2005; Knol & Berger, 1991). For additional information on approaches for assessing unidimensionality, see Hattie (1985) and Panter, Swygert, Dahlstrom, and Tanaka (1997).

NOHARM (Normal Ogive Harmonic Analysis Robust Method) is a general program that takes advantage of the relationship between nonlinear factor analysis and the normal ogive model to fit unidimensional and multidimensional normal ogive models (see Appendix C). In the current context, we fit one- and two-dimensional two-parameter (2P) models to the data. To determine which dimensional solution is “best,” the differences in fit among the models are examined. Because the models we are fitting do not address guessing, we are assuming the response data are not influenced by guessing. Moreover, we are also assuming the latent trait is normally distributed or is multivariate normal (McDonald, 1981). NOHARM is available as a stand-alone program (NOHARM4) or as a function within the R package *sirt* (Robitzsch, 2018). We use the R package *sirt* here; however, Endnote 3 shows the analysis using the stand-alone NOHARM4 program. (Appendix G, “R Introduction,” shows how to install R and obtain a package, and provides an introduction to setting up the R environment.) *sirt* is a flexible package for fitting dichotomous, polytomous, and multidimensional models as well as providing other useful functions (e.g., NOHARM).

Table 3.1 shows our R session. Our session begins with loading the *sirt* package (`library(sirt)`). We use the `read.table` function with the data filename specified (`Math.dat`) to read our data into our workspace and store it in the object (i.e., a data frame) `mathdata`. Because the data file does not contain variable names, we set `header=FALSE`. Consequently, R provides default names beginning with “V.” To verify that the data are read correctly, we display the first and last five cases of `mathdata` using the `head` and `tail` functions, respectively. Subsequently, we change the default variable names using the `names` function.

We proceed by fitting first a one-dimensional model and then a two-dimensional model. We invoke the `noharm` function by `noharm.sirt(mathdata, dimensions = 1, lower = 0, reliability = TRUE)` and store its output in the object `noharm1d`. Although not always necessary, it is usually efficient and convenient to store the output of a function in an object that can subsequently be used. For instance, we can indirectly access and/or use the object’s information by using it with other functions (e.g., `summary(modelfit.sirt(noharm1d))`) or directly accessed information by specifying the variable of interest. In this case, the syntax is “output object” + “\$” + variable name (e.g., `noharm1d$residuals`).

In our call to `noharm.sirt(...)`, we provide our data (`mathdata`) as the first argument. We specify (1) the fitting of a unidimensional model (`dimensions = 1`), (2) the setting of the IRF lower asymptotes for each item to 0 (`lower = 0`;

TABLE 3.1. sirt.noharm Session for Dimensionality Analysis^a

```

> library(sirt)
- sirt 3.4-64 (2019-05-03 18:33:11)

> mathdata=read.table("math.dat",header=FALSE)

> head(mathdat,n=5)    # show the first 5 cases of the data file
   V1 V2 V3 V4 V5
1  1  1  0  0  0
2  1  1  1  0  0
3  1  0  0  0  0
4  1  1  1  0  0
5  1  0  1  1  0

> # replace default variables names (i.e., V1,..., V5) with meaningful names
> names(mathdata)=c('i01','i02','i03','i04','i05')

> tail(mathdata,n=5)    # show the last 5 cases of the data file
   i1 i2 i3 i4 i5
19597 1  1  1  1  0
19598 1  1  1  1  0
19599 1  1  1  1  1
19600 1  1  0  1  1
19601 1  1  1  1  0

> # 1D analysis
> noharm1d=noharm.sirt(mathdata,dimensions=1,lower=0,optimizer="optim",
+                         reliability=TRUE)
> summary(noharm1d)
-----
sirt 3.4-64 (2019-05-03 18:33:11)
R version 3.6.0 (2019-04-26) i386, mingw32

Call:
noharm.sirt(dat = mathdata, dimensions = 1, lower = 0, optimizer = "optim",
             reliability = T)
:
Function 'noharm.sirt'
:
--- Information about optimization ---
Optimizer = optim
Converged = TRUE
Optimization Function Value = 5.4e-05
Number of iterations = 11
Elapsed time = Time difference of 0.01999998 secs

Number of Observations: 19601
Number of Items      : 5
Number of Dimensions : 1
Tanaka Index         : 0.99997
RMSR                : 0.00233                                ← The GFI
                                         ← The RMSR

Number of Used Item Pairs      : 10
Number of Estimated Parameters : 10
  # Thresholds            : 5
  # Loadings              : 5
  # Variances/Covariances : 0
  # Residual Correlations : 0

```

(continued)

TABLE 3.1. (continued)

```

Chi Square Statistic of Gessaroli & De Champlain (1996)

Chi2 : 81.238
Degrees of Freedom (df) : 5
p(Chi2,df) : 0
Chi2 / df : 16.248
RMSEA : 0.028 ← The RMSEA

Green-Yang Reliability Omega Total : 0.633

Factor Covariance Matrix
F1
F1 1
:
Factor Correlation Matrix
F1
F1 1
:
:

> summary(modelfit.sirt(noharm1d))
Test of Global Model Fit
    type   value      p
1  max(X2) 21.16443 4e-05
2 abs(fcor)  0.03507 0e+00

Fit Statistics
            est
MADcor      0.01013
SRMSR      0.01444 ← The SRMSR
100*MADRESIDCOV 0.17809
MADQ3       0.13653
MADAQ3     0.04681

> noharm1d$residuals
      i01        i02        i03        i04        i05
i01  0.000000e+00  0.0046288462 -7.818999e-05 -0.0040802357 -0.002610106
i02  4.628846e-03  0.0000000000 -8.768847e-04  0.0004925072 -0.002057442
i03 -7.818999e-05 -0.0008768847  0.000000e+00 -0.0001538747  0.001601404
i04 -4.080236e-03  0.0004925072 -1.538747e-04  0.0000000000  0.001246765
i05 -2.610106e-03 -0.0020574424  1.601404e-03  0.0012467653  0.000000000

> # 2D
> noharm2d=noharm.sirt(mathdata,dimensions=2,lower=0, optimizer="optim",
  reliability=T)
> summary(noharm2d)
-----
sirt 3.4-64 (2019-05-03 18:33:11)
R version 3.6.0 (2019-04-26) i386, mingw32
Call:
noharm.sirt(dat = mathdata, dimensions = 2, lower = 0, optimizer = "optim",
  reliability = T)
:
-- Information about optimization --
Optimizer = optim
Converged = TRUE

```

(continued)

TABLE 3.1. (continued)

```

Optimization Function Value = 1.1e-05
Number of iterations = 19
Elapsed time = Time difference of 0.03499985 secs

Number of Observations: 19601
Number of Items : 5
Number of Dimensions : 2
Tanaka Index : 0.99999 ← The GFI
RMSR : 0.00103 ← The RMSR

Number of Used Item Pairs : 10
Number of Estimated Parameters : 14
# Thresholds : 5
# Loadings : 9
# Variances/Covariances : 0
# Residual Correlations : 0

Chi Square Statistic of Gessaroli & De Champlain (1996)
Chi2 : 16.327
Degrees of Freedom (df) : 1
p(Chi2,df) : 0
Chi2 / df : 16.327
RMSEA : 0.028 ← The RMSEA

Green-Yang Reliability Omega Total : 0.646
:

> summary(modelfit.sirt(noharm2d))
Test of Global Model Fit
  type   value p
1  max(X2) 118.50265 0
2 abs(fcor)  0.07424 0

Fit Statistics
  est
MADcor      0.04624
SRMSR       0.05136 ← The SRMSR
100*MADRESIDCOV 1.04273
MADQ3        0.17190
MADAQ3       0.07028

> noharm2d$residuals
    i01          i02          i03          i04          i05
i01  0.000000e+00 -7.528227e-06  2.237367e-03 -2.111519e-03 -6.243539e-04
i02 -7.528227e-06  0.000000e+00 -6.472061e-04  5.537261e-04  1.621730e-04
i03  2.237367e-03 -6.472061e-04  0.000000e+00 -1.996900e-05 -6.026089e-05
i04 -2.111519e-03  5.537261e-04 -1.996900e-05  0.000000e+00  8.057347e-05
i05 -6.243539e-04  1.621730e-04 -6.026089e-05  8.057347e-05  0.000000e+00

```

^aThe '>' is the R prompt with the text following the prompt typed by the user. The '#' indicates a comment. We use '=' in lieu of the typical R assignment symbol of "<-" to be consistent with the syntax used in some programming languages (e.g., BASIC, C, C++, FORTRAN, Python, SAS), to reduce the number of characters, and to avoid potential problems such as typing 'x <-3' when 'x < -3' was intended; the tidy_source package can be used to replace '=' with '<-'.

i.e., we are assuming the response data are not influenced by guessing), and (3) the calculation of the reliability (`reliability = TRUE`). Our last argument specifies the use of the `optim` optimizer.⁴ To obtain our output, we use the `summary` function (`summary(noharm1d)`).

We first verify that the correct number of cases and items is used, that the dimensionality is what we want, and that we obtained a converged solution (see the “Information about optimization” section). As can be seen, the number of cases (Number of Observations) and items (Number of Items) are correct, as is the model’s dimensionality (Number of Dimensions). Our solution converged in 11 iterations (Converged = `TRUE` and Number of iterations = 11).

We use residuals to facilitate model–data fit analysis. The residual matrix is the discrepancy between the observed covariances and those predicted after the model has been fitted to the data (Fraser & McDonald, 2012). Thus, the ideal situation is where the discrepancies are zero. To summarize the residual matrix, NOHARM provides its root mean square residual (RMSR). The RMSR is the square root of the average squared difference between the observed and predicted covariances, with small values indicating a good fit. This overall measure of model–data misfit may be evaluated by comparing it to four times the reciprocal of the square root of the sample size (i.e., the “typical” standard error of the residuals; McDonald, 1997). Our RMSR of 0.00233 is less than the criterion ($4/\sqrt{19601} = 0.0286$) for these data.

A second useful measure is Tanaka’s (1993) goodness-of-fit index (GFI). McDonald (1999) suggests that a GFI of 0.90 indicates an acceptable level of fit and that a value of 0.95 indicates a “good” fit; GFI = 1 indicates perfect fit. Our GFI of 0.99997 indicates very good fit. Two more indices for assessing model–data fit are the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR). For RMSEA, values less than 0.05 are considered to indicate “close” fit. Our value is 0.028. With respect to SRMR, a value “close to” or less than 0.08 is thought to be good fit. Our SRMR of 0.01444 meets this guideline. Appendix G, “CFI, GFI, M_2 , RMSEA, TLI, and SRMR” contains additional information on GFI and SRMR.

To examine the residuals, we specify the output object with the residual variable (`noharm1d$residuals`). As can be seen, the residuals range from 0.00008 to 0.00463. Therefore, in light of the residuals and according to our indices, there does not appear to be sufficient evidence to reject a unidimensional solution.

The two-dimensional solution is performed by modifying the `dimensions` argument (`noharm.sirt(mathdata, dimensions=2, ...)`).⁵ As is the case with the one-dimensional solution, the two-dimensional solution’s residuals are comparatively small and the matrix does not reveal any large residuals. Not surprisingly, as the dimensionality of the models increased, the corresponding residuals decreased and, therefore, so did the RMSR. The solution’s RMSR of 0.00103 is substantially less than the criterion of 0.0286. With respect to Tanaka’s index, the two-dimensional solution’s value is 0.99999. The RMSEA and SRMR values are 0.028 and 0.05136, respectively, and indicate good fit. Although the two-dimensional 2P model has the lower RMSR and larger Tanaka index, the application of Occam’s razor leads us to *not* reject the unidimensional model of the data. Therefore, we conclude that our unidimensional model is a sufficiently

accurate representation of the data to proceed with the IRT calibration.⁶ (Appendix G, “An Approximate Chi-Square Statistic for NOHARM,” discusses the Chi Square Statistic of Gessaroli and De Champlain.)

Calibration Result, BIGSTEPS

With BIGSTEPS an ASCII (i.e., text) input file is created that contains control information for the calibration. The first few lines of this file are presented below; the text following an “←” is not part of the file and is an annotation.

```
;Rasch Calibration of Mathematics data
&INST      ← required (NAMELIST structure, start)
TITLE='rasch calibration math data'
NI=5       ← number of items
ITEM1=6    ← item responses begin in column 6
XWIDE=1    ← item responses occupy 1 column
CODES=01   ← possible item response codes; responses are binary
NCOLS=10   ← last column of item responses
MODELS=R   ← model declaration
STBIAS=Y   ← use bias correction factor7
GROUPS=0   ← part of model declaration
TABLES=1110011001101000100000 ← specifying which output
                                tables to produce
:
&END      ← required (NAMELIST structure, end)
:
```

For this example, a single file (MATH.CON) contains both the control information and the response data. Alternatively, we can keep the control information and the data in separate files. If the data reside in a separate file, our command file would need the DATA command followed by the data file name (e.g., DATA=MATH.DAT; it is assumed that all files reside in the same folder [i.e., subdirectory]). The number of items is identified with the NI command. The ITEM1 and XWIDE commands are used to define how the response data are read. For example, our ITEM1 and XWIDE commands indicate that the response vectors begin in column 6 and that each response occupies one column, respectively. Because these data are already in a binary format, the CODES command simply identifies the values 0 and 1. However, if the data are not in a 0/1 format, then the response codes in the data file would be identified on the CODES command and a KEY1 command would be inserted into the command file to identify which responses should be converted to 1; by omission, all other responses are converted to 0.

BIGSTEPS can produce a plethora of results. All results are presented in tables, and the user can control which tables are presented in the output file by using the TABLES command. The TABLES command consists of 1s and 0s, where a 1 indicates that a table should be displayed in the output file; otherwise it should be suppressed. For instance, our TABLES command indicates that we want tables 1, 2, and 3 displayed in the output file, tables 4 and 5 should be suppressed, tables 6 and 7 should be shown in the output file, and so on.

The output file always contains two tables. These default tables, TABLE 0.1 and

TABLE 0.2, appear at the *end* of the output file and are presented in Table 3.2. (We use the Courier font for BIGSTEPS's table labels.) Table 0.1 shows the program control parameter settings (e.g., convergence criterion, number of items, metric scaling), whereas TABLE 0.2 presents the iteration history. These tables should be the first tables examined. TABLE 0.1 should be inspected to verify that the program control parameters are correct and the data have been read correctly. Subsequently, the iteration history (TABLE 0.2) should be examined to ensure that convergence is achieved.

BIGSTEPS (and WINSTEPS) uses a normal approximation estimation procedure called PROX as a preprocessing step to provide good starting values for the UCON (JMLE) procedure; PROX assumes a normal distribution of ability to simplify calculations. We see that PROX iterated three times before achieving convergence; TABLE 0.1 shows that the default maximum iterations for PROX (i.e., MPROX) is 10. The ACTIVE COUNT columns indicate how many persons, items, and categories (5 items * two categories: 0, 1) the program started with and how many remain after deletion of zero and perfect response vectors (i.e., response vectors that contain all 0s or all 1s, respectively). After removing 4076 people (19,601 – 15,525) that had either 0 or perfect response vectors, there are 15,525 examinees remaining.⁸ If the number of CASES under ACTIVE COUNT at the first iteration does not match the number of persons to be used in the calibration, then the data file was not correctly read. In this example, the ACTIVE COUNT CASES column shows the program read the correct number of individuals (i.e., 19,601). The EXTREME 5 RANGE columns provide an estimate of the dispersion between the mean $\hat{\theta}$ of the top 5 and the mean $\hat{\theta}$ of the bottom 5 persons (labeled CASES) as well as a current estimate of the spread between the mean $\hat{\delta}$ of the top 5 and the mean $\hat{\delta}$ of the bottom 5 (labeled ITEMS). MAX LOGIT CHANGE is the logit change across iterations for persons (labeled MEASURES) and, if relevant, for items (labeled STEPS). In a well-behaved situation, these values should decrease as the number of iterations increases.

The bottom portion of Table 0.2 contains the iteration history for UCON. By default the maximum number of UCON iterations is unlimited, but we have it set to 25 (see TABLE 0.1: MUCON = 25) to avoid having to abort the program's execution in case of convergence problems. That is, if UCON performs a large number of iterations, then it is having difficulty converging to a solution. This may be due to many reasons, such as the data not conforming to the model or to a misspecified command file (e.g., the item responses are not in the specified columns). In this example, UCON converged in 17 iterations. Convergence for UCON is controlled by criteria associated with MAX SCORE RESIDUAL and MAX LOGIT CHANGE columns. The MAX SCORE RESIDUAL is the maximum change in the residuals for an item or a person across iterations (i.e., the numerator of the step size discussed in Appendix A); the SCORE RESIDUAL is the difference between observed and expected scores (for a person or an item). In contrast, MAX LOGIT CHANGE is the maximum change in the location of a person or an item from one iteration to the next. Both of these should progressively decrease as the program iterates to a solution. Convergence is defined by RCONV for MAX SCORE RESIDUAL and by LCONV for MAX LOGIT CHANGE. In this example, the (default) convergence criterion for items is 0.01 (see TABLE 0.1: LCONV = 0.010) and for persons it is 0.5

TABLE 3.2. BIGSTEPS Abridged Program Control Parameters (Table 0.1) and Iteration History Table (Table 0.2)

TABLE 0.1 rasch calibration math data

```

TITLE= rasch calibration math data
CONTROL FILE: MATH.CON
OUTPUT FILE: MATH.LST

CONTROL VARIABLES:
Input Data Format      PAIRED = N           Item Delete/Anchor
  DATA =                REALSE = N          IDFILE =
  NAME1 = 1              STBIAS = Y          IDELQU =
NAMLEN = 5             -----          IAFILE =
ITEM1 = 6               Misfit Selection   IANCHQ = N
ITLEN = 30              FITI = 2.000       -----
NI = 5                 FITP = 2.000       Person Delete/Anchor
XWIDE = 1               OUTFIT = Y        PDFILE =
:
:
PERSON = CASE          -----
ASCII = Y              Convergence Control RFILE =
-----          MPROX = 10          SFILE =
MUCON = 25            LCONV = .010        XFILE =
User Scaling          RCONV = .500       -----
UMEAN = .000           TARGET = N        Data Reformat
USCALE = 1.000         GROUPS = 0        FORMAT =
UDECIM = 2             MODELS = R        GRPFRM = N
UANCH = Y              STKEEP = N        KEYFRM = 0
-----          Scale Structure    MODFRM = N
Adjustment             RESFRM = N        SPFILER =
EXTRSC = .500          -----
HIADJ = .250           -----
LOWADJ = .250          -----
19601 CASE  Records Input

```

TABLE 0.2 rasch calibration math data
INPUT: 19601 CASES, 5 ITEMS

CONVERGENCE TABLE

PROX ITERATION	CASES	ACTIVE COUNT ITEMS	CATS	EXTREME CASES	5 RANGE ITEMS	MAX LOGIT MEASURES	CHANGE STEPS
1	19601	5	10	2.77	1.14	2.0650	
2	15525	5	10	3.17	1.69	.8874	
3	15525	5	10	3.58	1.76	.2125	
UCON ITERATION	MAX SCORE RESIDUAL*	MAX LOGIT CHANGE	LEAST CASE	CONVERGED ITEM	CATEGORY CAT	CATEGORY RESIDUAL	STEP CHANGE
1	428.35	-.1716	2	5*			
2	153.49	-.0581	2	5*			
3	92.02	-.0434	2	4*			
4	58.32	-.0317	7	4*			
5	39.53	-.0211	7	4*			
6	26.04	-.0147	167	4*			
7	16.26	-.0097	167	4*			
8	11.04	-.0063	167	4*			
9	5.72	-.0043	167	5*			
10	4.40	-.0023	7	4*			
11	2.74	-.0016	167	4*			
12	1.24	-.0010	167	4*			
13	-1.24	-.0006	167	3*			
14	.94	-.0006	167	4*			
15	-.66	-.0004	167	2*			
16	-.65	-.0003	167	2*			
17	-.49	-.0002	167	3*			

Standardized Residuals N(0,1) Mean: .01 S.D.: 1.03

(see TABLE 0.1: RCONV = 0.500). We can see that both MAX SCORE RESIDUAL and MAX LOGIT CHANGE decrease as UCON iterates to convergence.

The LEAST CONVERGED columns present for persons (labeled CASE) the ordinal position of the person that is farthest from meeting convergence criterion and for items (labeled ITEM) the ordinal position of the item that is farthest from meeting the criterion. For persons (CASE) individuals 2, 7, 167 were identified, and for items (ITEM) 2 through 5 were flagged. If we had experienced a convergence problem, then this information might be useful in rectifying the problem.

At the bottom of Table 0.2, we find the MEAN and standard deviation (S.D.) of the standardized residuals. These residuals should follow a normal distribution and have a mean close to 0.0 and a standard deviation of approximately 1.0. MEAN and S.D. values that are substantially different from these ideal values indicate the data do not follow the Rasch model assumptions (Linacre & Wright, 2001). Moreover, if these descriptive statistics indicate a departure from the assumption that randomness is normally distributed, then the fit statistics, INFIT and OUTFIT, are adversely affected. Our descriptive statistics do not indicate any problems. Because TABLE 0.2's information does not show any difficulties in calibrating the data, we proceed to examine model-level fit and then item-level fit.

Model-Data Fit Assessment

Model-Level Fit

BIGSTEPS's TABLE 3.1 (see Table 3.3) contains model-level fit results. TABLE 3.1 is divided into a top half containing respondent information (labeled SUMMARY OF ... CASES) and a bottom half containing item information (labeled SUMMARY OF ... ITEMS).

TABLE 3.1's second line (labeled INPUT) indicates that 15,525 of 19,601 examinees and all five items were used for calibration (i.e., 4076 examinees were dropped, but no items were removed from the calibration). These 15,525 respondents are the NON EXTREME examinees. (In BIGSTEPS parlance, respondents are classified as either EXTREME or NON EXTREME. An EXTREME respondent has either a zero or perfect observed score, whereas a NON EXTREME respondent does not.) From the top half of TABLE 3.1 we see why 4076 individuals were removed from the analysis. Specifically, 3385 of the 4076 dropped examinees obtained the MAXIMUM EXTREME SCORE of 5 (i.e., these persons correctly responded to all items, $X = 5$), and 691 obtained the MINIMUM EXTREME SCORE of 0 (i.e., these persons incorrectly responded to all items).

The MEAN observed score for the 15,525 individuals is 2.6 with an S.D. of 1.1. After removing the zero and perfect scores, the maximum (MAX.) and minimum (MIN.) observed scores are 4 and 1, respectively. The COUNT column refers to the number of items attempted. The COUNT column's MEAN and S.D. of 5 and 0, respectively, reflect that no items were omitted. These descriptive statistics are also presented in logits for both person $\hat{\theta}$ s (labeled MEASURE) and their corresponding standard errors (labeled MODEL ERROR). The average person location estimate is 0.18 with an average standard

TABLE 3.3. Rasch (BIGSTEPS) Calibration Summary Results for the Mathematics Test Data

SUMMARY OF 15525 MEASURED (NON-EXTREME) CASES								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	2.6	5.0	.18	1.19	.98	.0	1.03	.1
S.D.	1.1	.0	1.39	.13	.66	1.0	1.40	1.2
MAX.	4.0	5.0	1.95	1.42	3.37	4.5	9.90	9.8
MIN.	1.0	5.0	-2.00	1.06	.28	-1.3	.16	-.7
REAL RMSE	1.34	ADJ.SD	.39	SEPARATION	.29	CASE	RELIABILITY	.08
MODEL RMSE	1.19	ADJ.SD	.72	SEPARATION	.61	CASE	RELIABILITY	.27
S.E. OF CASE MEAN	.01							
WITH 4076 EXTREME CASES	= 19601 CASES		MEAN	.53	S.D.	1.75		
REAL RMSE	1.39	ADJ.SD	1.06	SEPARATION	.76	CASE	RELIABILITY	.37
MODEL RMSE	1.28	ADJ.SD	1.19	SEPARATION	.93	CASE	RELIABILITY	.46

SUMMARY OF 5 MEASURED ITEMS								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	8029.8	15525.0	.00	.02	1.00	-1.2	1.06	-.2
S.D.	3499.1	.0	1.24	.00	.11	7.3	.21	7.2
MAX.	14010.0	15525.0	1.27	.02	1.17	9.9	1.39	9.9
MIN.	4207.0	15525.0	-2.21	.02	.86	-9.9	.78	-9.9
REAL RMSE	.02	ADJ.SD	1.24	SEPARATION	68.25	ITEM	RELIABILITY	1.00
MODEL RMSE	.02	ADJ.SD	1.24	SEPARATION	70.09	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN	.62							

error of 1.19; the standard deviation of the person location estimates is 1.39. The least able person is located at -2 (i.e., MIN.), and the most able person is at 1.95 (i.e., MAX.).

BIGSTEPS (and WINSTEPS) produce two fit statistics, INFIT and OUTFIT, for examining model-data fit. In the following paragraphs, we define these statistics and then discuss them in the context of this example. INFIT and OUTFIT provide information concerning discrepancies in responses, depending on whether the discrepancies occur close to or farther away from the estimated parameter. These fit statistics are calculated for both persons and items at the individual person and item levels. Therefore, each individual and item has both an INFIT and OUTFIT statistic. For persons the statistics' calculations involve a sum across items, whereas for items the sum is across people.

INFIT is a weighted fit statistic based on the squared standardized residual between what is observed and what would be expected on the basis of the model (i.e., a chi-square-like statistic). These squared standardized residuals are information weighted and then summed across observations (i.e., items or people); the weight is $p_j(1 - p_j)$.

These “chi-square” statistics are averaged to produce the `INFIT` mean-square statistic; the mean square is labeled `MNSQ` in the output.

`OUTFIT` is also based on the squared standardized residual between what is observed and what would be expected, but the squared standardized residual is not weighted when summed (i.e., across items or people, whichever is relevant). As such, `OUTFIT` is an *unweighted* standardized fit statistic. As is the case with the `INFIT` statistic, the `OUTFIT` statistic is transformed to a mean square and labeled `MNSQ` in the output.

These two statistics differ in their sensitivity to where the discrepancy between what is observed and what is expected occurs. For instance, and from a person fit perspective, responses on items located near the person’s $\hat{\theta}$ that are in line with what would be expected produce `INFIT` values close to 1 (given the stochastic nature of the model). However, responses on items located near the person’s $\hat{\theta}$ that are not in line with what would be expected lead to large `INFIT` values. That is, `INFIT` is sensitive to unexpected responses near the person’s $\hat{\theta}$. In contrast, `OUTFIT` has a value close to its expected value of 1 when responses on items located away from a person’s $\hat{\theta}$ are consistent with what is predicted by the model (again, given the stochastic nature of the model). However, unexpected responses on items located away from a person’s $\hat{\theta}$ (i.e., “outlier” responses) lead to `OUTFIT` values substantially greater than 1. That is, `OUTFIT` is sensitive to, say, a high-ability person incorrectly responding to an easy item or a low-ability person correctly responding to a hard item. One has an analogous interpretation for these fit statistics when used for item fit analysis.

The range of `INFIT` and `OUTFIT` is 0 to infinity with an expectation of 1; their distributions are positively skewed. Values that are above or below 1 indicate different types of misfit. For example, values substantially less than 1 may be indicative of dependency or overfit, whereas values substantially greater than 1 may reflect noise in the data. Although there are various interpretation guidelines, one guideline states that values from 0.5 to 1.5 are “okay,” with values greater than 2 warranting closer inspection of the associated person or item. Smith, Schumacker, and Bush (1998) state that using a common cutoff value does not necessarily result in correct Type I error rates. They echo Wright’s suggestion (see Smith et al., 1998) to take sample size into account when interpreting `INFIT` and `OUTFIT` by using $1 \pm 2/\sqrt{N}$ and $1 \pm 6/\sqrt{N}$ as cutoff values, respectively.

Given `INFIT`’s and `OUTFIT`’s expectations and their range, it is clear there is an asymmetry in their scales. Therefore, `INFIT` and `OUTFIT` are transformed to have a scale that is symmetric about 0.0. This transformation results in a standardized fit statistic, `ZSTD`; `ZSTD`s are obtained by using a cube root transformation of the `MNSQ`s to make them “normally” distributed and to have a range from $-\infty$ to ∞ . Good fit is indicated by `INFIT ZSTD` and `OUTFIT ZSTD` values close to 0. Because the `ZSTD`s are approximate *t* statistics, as sample size increases these “*t* statistics” approach *z* statistics. As such, values of ± 2 are sometimes used for identifying items or people that warrant further inspection; for inferential testing, the null hypothesis is perfect model–data fit. In our output, the standardized `INFIT` statistic is labeled `INFIT ZSTD`, and the standardized `OUTFIT` statistic is labeled `OUTFIT ZSTD`. See Linacre and Wright (2001) and Smith (1991, 2004) for more information on `INFIT` and `OUTFIT` and their transformations.

Returning to our example, one sees that the top half of BIGSTEPS's TABLE 3.1 (see Table 3.3) contains descriptive statistics for the *overall* person fit, whereas the bottom half contains descriptive statistics associated with *overall* item fit. We begin with the MNSQs. Because the mean INFIT MNSQ and OUTFIT MNSQ values are close to their expected value of 1.0, we conclude that most of the participants are behaving consistently with the model. The variability of these fit statistics (S.D. = 0.66 for INFIT and S.D. = 1.40 for OUTFIT) as well as the maximum fit statistics (e.g., maximum OUTFIT MNSQ = 9.9) indicate that not all individuals are responding consistently with the model. The reasons for these maxima are discussed below.

Additional fit information provided in TABLE 3.1 comes from REAL RMSE and MODEL RMSE. REAL RMSE and MODEL RMSE are calculated both with and without deleted persons (i.e., with and without the 4076 EXTREME CASES). REAL RMSE is the root mean squared error calculated from the perspective that misfit is due to departures in the data from model specifications, whereas MODEL RMSE is calculated from the perspective that the data fit the model; "REAL" means the statistics have been adjusted for misfit encountered in the data (Linacre & Wright, 2001). Both RMSEs are calculated for persons and items. The top half of TABLE 3.1 contains the calculations for people, and the bottom half contains the calculations for items. Small values of these two statistics indicate a good situation, with large values of REAL RMSE reflecting departures in the data from the model. These statistics' values indicate that we're doing a better job with the items than with the people.

In general, one typically administers an instrument to differentiate between persons located at different points along the latent variable. However, we may also be interested in how well an instrument can separate or distinguish items in terms of their latent variable locations. In this regard, the program produces the SEPARATION index for persons and for items.

The person SEPARATION index gives an indication of how well the instrument can separate or distinguish persons in terms of their latent variable locations. Although this index has a lower bound of 0, it does not have an upper bound. Because the SEPARATION index does not have a finite upper bound, it is sometimes difficult to determine what a good large value is. In contrast, the related RELIABILITY index is more easily interpreted than the SEPARATION index. Similar to coefficient alpha Guttman (1945) λ_3 , the person RELIABILITY tells us about the consistency or reproducibility of the $\hat{\theta}$ s. Its range is from 0 to 1, with values close to or at 1 considered better than lower values.⁹ For our example, the RELIABILITY for the REAL RMSE (non-extreme) line is 0.08. This value indicates that the mathematics instrument does not appear to be doing a good job of distinguishing people. We conjecture that this is due, in part, to the instrument's length of five items. These indices are discussed further in Appendix G, "The Separation and Reliability Indices."

In the subsection entitled "WITH . . . EXTREME CASES . . ." the RMSEs, ADJ.SDs, SEPARATIONS, and RELIABILITYs are repeated using all 19,601 respondents (i.e., including the 4076 EXTREME persons). Some may find comparing these results with those from the NON EXTREME individuals useful in determining

the effect of the EXTREME persons on the overall fit. However, we ignore these results because they are affected by the value used in the estimation of the EXTREME persons (i.e., the EXTRSC estimation adjustment criterion that is discussed below).

The bottom half of TABLE 3.1 (labeled SUMMARY OF 5 MEASURED ITEMS) presents analogous information for items; MEASURE in this section refers to $\hat{\delta}$. As is the case with the person half, descriptive statistics for the items, both in raw score and logit units, are provided. For example, the mean item location estimate in logits is 0.0, with a standard deviation (labeled S.D.) of 1.24. It can also be seen that the lowest item location estimate is -2.21 (labeled MIN.) and the highest is located at 1.27 (labeled MAX.). Comparing these minimum and maximum location estimates with those of the persons (minimum = -2 and maximum = 1.95) shows a match at the lower end of the continuum between the instrument's $\hat{\delta}$ and the minimum persons location estimate, but not at the upper end. Overall, the INFIT/OUTFIT MNSQ values, their lack of variability, and their minima and maxima indicate the instrument *as a whole* appears to have model-data fit. This does not necessarily mean that we have fit for each item. Therefore, in the next section we examine item-level fit.

Analogous to the person SEPARATION index, the item SEPARATION index indicates how well the instrument can separate or distinguish items in terms of their latent variable locations. The premise of this index is that we would like our items to be sufficiently well separated in terms of their locations in order to identify the direction and meaning of the latent variable (Wright & Masters, 1982). In addition, we would like to see little estimation error. This last aspect is assessed by the (item) RELIABILITY index. As described above with the person SEPARATION and RELIABILITY indices, our item SEPARATION and RELIABILITY indices are related in an analogous fashion. Therefore, despite the lack of an upper bound for the item SEPARATION index, we can interpret the item RELIABILITY index. Our REAL ITEM RELIABILITY of 1.00 indicates the instrument is creating a well-defined variable, although it is apparently not performing well with respect to these people. It should be noted that whether this instrument is measuring the intended latent variable requires a validity study.

Item-Level Fit and Item-Location Estimates

Table 3.4 contains two BIGSTEPS tables, TABLE 13.1 and TABLE 13.2. We examine the item information presented in Table 3.4 prior to the person information because the ratio of persons to items favors item estimation. If there are problems in estimating the items, then these problems could potentially affect the estimation of the persons. Therefore, there would have been little reason to examine the person estimation output.

The column in TABLE 13.1 labeled MEASURE contains the item location estimates, $\hat{\delta}$ s, and the ERROR column contains the corresponding standard errors of estimate, $s_e(\hat{\delta})$ s. The column labeled RAW SCORE is the observed item score (i.e., the number of correct responses on item j , q_j), and COUNT is the number of respondents to the item; RAW SCORE divided by COUNT gives the item's traditional item difficulty, P_j . This table is displayed in descending order of item difficulty (i.e., MEASURE ORDER).

TABLE 3.4. Rasch (BIGSTEPS) Item Location Estimates for the Mathematics Test Data

TABLE 13.1 rasch calibration math data
 INPUT: 19601 CASES, 5 ITEMS ANALYZED: 15525 CASES, 5 ITEMS, 10 CATS v2.82

ITEMS STATISTICS: MEASURE ORDER

ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTBIS CORR.	ITEMS G
					MNSQ	ZSTD	MNSQ	ZSTD		
5	4207	15525	1.27	.02	1.17	9.9	1.39	9.9	-.07	ITEM5 0
4	4984	15525	1.01	.02	.98	-2.3	1.05	2.4	.08	ITEM4 0
3	7709	15525	.20	.02	.93	-7.4	.91	-7.0	.15	ITEM3 0
2	9239	15525	-.26	.02	.86	-9.9	.78	-9.9	.22	ITEM2 0
1	14010	15525	-2.21	.02	1.07	3.7	1.17	3.4	.02	ITEM1 0
MEAN	8030.15525.		.00	.02	1.00	-1.2	1.06	-.2		
S.D.	3499.	0.	1.24	.00	.11	7.3	.21	7.2		

:

TABLE 13.2 rasch calibration math data
 INPUT: 19601 CASES, 5 ITEMS ANALYZED: 15525 CASES, 5 ITEMS, 10 CATS v2.82

ITEMS FIT GRAPH: MEASURE ORDER

ENTRY NUMBR	MEASURE	INFIT MEAN-SQUARE				OUTFIT MEAN-SQUARE				ITEMS G
		-	+	0	0.7 1 1.3	2	0	0.7 1 1.3	2	
5	*	:	.	*	:		:	.	*	ITEM5 0
4	*	:	*	.	:		:	*	:	ITEM4 0
3	*	:	*	.	:		:	*	.	ITEM3 0
2	*	:	*	.	:		*	.	:	ITEM2 0
1	*	:	*	.	:		:	*	:	ITEM1 0

:

The item location estimates for our items are $\hat{\delta}_1 = -2.21$, $\hat{\delta}_2 = -0.26$, $\hat{\delta}_3 = 0.20$, $\hat{\delta}_4 = 1.01$, and $\hat{\delta}_5 = 1.27$, each with a standard error of 0.02.

TABLE 13.1 also shows the values of the item-level fit statistics in the INFIT MNSQ and OUTFIT MNSQ columns. Recall that values around 1 are considered good, with values substantially different from 1 indicating either dependency or noise. Following Smith et al.'s (1998) suggestion of using $1 \pm 2/\sqrt{N}$ and $1 \pm 6/\sqrt{N}$ for defining INFIT and OUTFIT screening values, the acceptable range for INFIT would be 0.9839–1.0161 (inclusive) and for OUTFIT the range would be 0.9518–1.0482 (inclusive). Using these criteria, and regardless of whether INFIT or OUTFIT is examined, items 1 and 5 would be flagged as “exhibiting misfit” and warranting further inspection. However, with a sample size of 15,525, these criteria result in very little tolerance of deviating from the expected value of 1. (There does not appear to be an empirical justification for the sample size-dependent criteria.)

The direct relationship between sample size and power is well known. Thus, given that our $N = 15,525$, we choose to ignore the INFIT and OUTFIT statistical tests ZSTD

(see Linacre, 2003). With a substantially smaller sample size, we would use the statistical interpretation ZSTDs to complement the MNSQs.

Additional item fit information is available in TABLE 13.2's "ITEMS FIT GRAPH." In this graph, the vertical lines consisting of colons delimit common cutoff points at 0.7 and 1.3. (This guideline is for a "run of the mill" instrument [Wright & Linacre, 1994]. However, for "high-stakes" and rating scale data, they suggest the guideline is 0.8–1.2 and 0.6–1.4, respectively.) This graph shows that all the items essentially fall within the 1 ± 0.3 interval. That is, there does not appear to be any apparent serious fit problems because the fit statistics are within acceptable bounds. (The use of a common cutoff ignores the fact that large samples tend to provide more accurate estimates than smaller samples, all things being equal. Moreover, the corresponding statistical test's Type I error rate is not maintained at the significance level when using a common cutoff [Smith et al., 1998]. As such, some individuals do not consider these common cutoffs as useful as criteria that take into account sample size, like those in the paragraph above.) We view the INFIT and OUTFIT guidelines not as hard-and-fast, but as providing guidance for identifying items and people that warrant closer inspection.

Although we feel comfortable with the above evidence supporting item-level data fit, our discussion continues in order to show one approach for diagnosing misfit. In practice, any item(s) flagged as potentially problematic are examined in an attempt to determine the cause of the misfit. For example, is the misfit due to misfitting people, is it due to individuals located far away from the item's location not responding as expected, is the misfit due to a few individuals or many, or is it due to poor item wording? Depending on the diagnosis, we may need to consider eliminating the item from the instrument. In this situation and if we have multiple candidate items, we would remove the very worst misfitting item(s) before less worse misfitting item(s). If we eliminate one or more items, then it would be necessary to recalibrate the instrument, because each individual could have a potentially different observed score and the original estimate of the person's location could have been adversely affected by the deleted item(s). After recalibration, we would need to reexamine our model–data fit. Potentially, there may be more than one additional calibration iteration because, for example, one or more misfitting items in the initial calibration masked another misfitting item. In these cases, one should (as part of the model–data fit analysis) also compare the estimates across the iterations to determine the magnitude of the impact of the misfitting item(s) on the estimates. If the impact is "minimal," then retaining the misfitting item(s) may be beneficial in a validity study; what is considered "minimal" is context specific (e.g., high-stakes testing, attitude scale).

As an example of following up on misfitting items, we will use the items flagged using Smith et al.'s (1998) guidelines. The items flagged for further inspection are the easiest (item 1) and hardest (item 5) items on the instrument. This may be related to the above-mentioned maximum person OUTFIT MNSQ value of 9.9 (OUTFIT ZSTD = 9.8). For example (and recalling that the items are ordered by difficulty), the response patterns 00001 and 01111 would lead to misfitting persons. In fact, Table 2.1 (Chapter 2) shows there are 184 persons with a pattern of 00001 and 40 with a pattern of 01111. For the 00001 pattern, the most difficult item is correctly answered (e.g., by guessing) and all

easier items are not. This pattern is inconsistent with the model. Conversely, the 01111 pattern shows the easiest item is incorrectly answered (e.g., owing to inattentiveness), but the more difficult items are answered correctly. Again, this pattern is inconsistent with the model. Individuals providing either of these two patterns would most likely be identified as misfitting. As such, person misfit can impact item misfit.

In its TABLE 11.1, BIGSTEPS provides an item's response vector (i.e., all the responses to an item) with the corresponding standardized residuals to aid in diagnosing item misfit. Table 3.5 contains a small portion of TABLE 11.1. TABLE 11.1's format is the item label (e.g., for item 5 the label is 5 ITEM5; "ITEM5" is a user-supplied

TABLE 3.5. Abridged Item Misfit Information for Poorly Fitting Items 1 and 5

label) along with the item's location estimate ($\hat{\delta}_5 = 1.27$) and its INFIT ZSTD and OUTFIT ZSTD values. The subsequent line contains the responses (RESPONSE) from the first 25 persons to item 5 (i.e., persons 1–8 responded incorrectly to item 5, person 9 correctly responded, person 10 responded incorrectly, and so on). The next line shows the standardized residuals (Z-RESIDUAL) information corresponding to each of these responses. A positive number indicates an unexpected response from primarily lower-ability individuals (i.e., a correct response), and a negative value reflects an unexpected response from primarily higher-ability persons (i.e., an incorrect response); an X reflects that the person providing the response is an EXTREME individual. These latter two lines are repeated for the remaining persons in the data set.

We see from Table 3.5 that there are a number of standardized residuals of 2 and 5 for item 5, as well as -3, -5, and -9 for item 1. Matching the response data with the persons with Z-RESIDUALs of 5 shows they come from individuals (persons 4,815 and 13,500) who provided the response vector 00001; the Z-RESIDUAL = 2 comes from individuals (persons 25, 38, etc.) with the response vector 10001. With respect to item 1, the Z-RESIDUAL of -9 (person 358) is associated with a response vector of 01111 (i.e., an incorrect response to the easiest item from a high-ability person), a Z-RESIDUAL of -5 (person 45) is associated with $\underline{x} = 01110$, and Z-RESIDUAL of -3 (person 20) comes from $\underline{x} = 01010$. In short, the largest absolute standardized residuals come from persons with response vectors that are inconsistent with the model. Using just this information (i.e., as opposed to the question's wording), we attribute items 1 and 5's fit issue to these individuals and do not consider fit to be an issue. These individuals are discussed below when we review individual person fit information.

Person Information

Table 3.6 (abridged TABLE 17.1) shows the person location estimates and corresponding fit information. ENTRY NUMBER is the case's ordinal position in the file, RAW SCORE is the person's observed score (X), and COUNT is the number of items answered, MEASURE is the person's location (proficiency) estimate with its corresponding standard error of estimate (ERROR). Only the unique $\hat{\theta}$ s are shown in Table 3.6. For observed scores of 0, 1, 2, 3, 4, and 5, the corresponding person estimates, $\hat{\theta}$ s, are -3.19 (1.72), -2.00 (1.42), -0.43 (1.13), 0.73 (1.06), 1.95 (1.21), and 2.86 (1.55), respectively; the associated $s_e(\hat{\theta})$ s are in parentheses. For instance, all individuals who got one item correct (i.e., $X = 1$) have an estimated mathematics proficiency of -2.00, all individuals who correctly answered two items have -0.43 as their person location estimate, and so on (i.e., every person with a given X receives the same $\hat{\theta}$).

It may have been noted that BIGSTEPS provided $\hat{\theta}$ s for $X = 0$ and $X = 5$ despite the fact that in Chapter 2 we demonstrated that it is not possible to obtain finite estimates for zero or perfect scores using MLE. BIGSTEPS obtains these location "estimates" by using a kludge. Specifically, it subtracts a fractional score point from the perfect score and adds the same fractional score point to the 0 score to create scores that are estimable. By default, this fractional score point is 0.5 (the control parameter EXTRSC can be used to change this value).¹⁰ Therefore, for estimating these EXTREME persons the perfect

TABLE 3.6. Abridged Person Estimate Table

TABLE 17.1 rasch calibration math.data INPUT: 19601 CASES, 5 ITEMS ANALYZED: 15525 CASES, 5 ITEMS, 10 CATS v2.82											
CASE STATISTICS: MEASURE ORDER											
ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTBIS CORR.	CASE	
9	5	5	2.86	1.55	MAXIMUM ESTIMATED MEASURE				9		
:											
	7	4	5	1.95	1.21	.72	-.5	.45	-.2	.55	7
:											
	2	3	5	.73	1.06	.49	-1.3	.40	-.5	.80	2
:											
	1	2	5	-.43	1.13	.55	-.8	.41	-.7	.84	1
:											
	3	1	5	-2.00	1.42	.28	-.8	.16	-.5	.85	3
:											
19414	0	5	-3.19	1.72	MINIMUM ESTIMATED MEASURE				19414		
19585	0	5	-3.19	1.72	MINIMUM ESTIMATED MEASURE				19585		
MEAN	3.	5.	.18	1.19	.98	.0	1.03	.1			
S.D.	1.	0.	1.39	.13	.66	1.0	1.40	1.2			

score of 5 is converted to 4.5, and the zero score becomes 0.5. Hence, the $\hat{\theta}$ of -3.19 is *not* the MLE corresponding to zero score, and the $\hat{\theta}$ of 2.86 is *not* the MLE for a perfect score. Rather, these are artifacts of the particular fractional score point value used; some may consider this modification of X to be a form of imputation. It should be recalled these extreme persons (i.e., $X = 0$ and $X = 5$) are not used for estimating item locations. Therefore, the choice of the fractional value does not affect the item location estimates.

The individual person INFIT and OUTFIT (Table 3.6) are defined analogously to the way they are defined for items (cf. Linacre & Wright, 2001, pp. 92–93). We do not present this fit information for all persons because there are potentially many INFIT and OUTFIT values across our 19,601 examinees. That is, even though all persons with the same observed score have the same estimated location (and error), this does not mean the fit is the same for all persons with the same observed score. Rather, persons with the same response pattern have the same INFIT and OUTFIT values. For instance, persons with the pattern 11110 (i.e., a pattern consistent with the model) have INFIT and OUTFIT values of 0.72 and 0.45, respectively. However, persons with the same observed score, but with the pattern 11011, have an INFIT value of 1.46 and an OUTFIT value of 1.36. The interpretation of INFIT and OUTFIT is similar to that used with the

item-level analysis. The overall model-level person fit information presented in Table 3.3 is based on the individual-level fit statistics shown in Table 3.6. As shown below, this individual-level person fit information facilitates identifying person-level problems.

Above it is mentioned that BIGSTEPS's TABLE 3.1 shows a maximum person OUTFIT MNSQ value of 9.90 (OUTFIT ZSTD = 9.8). We can explore this result by using different BIGSTEPS tables. For instance, BIGSTEPS can provide tables containing a plot of each individual's INFIT and OUTFIT values against their $\hat{\theta}$; these are BIGSTEPS's Tables 4 and 5. These plots graphically depict the extensiveness of person misfit and the point(s) (if any) along the continuum where misfit is occurring, and they also give a sense of the distribution of these fit statistics. In addition, BIGSTEPS provides a table (TABLE 7.1) that displays the poorly fitting individuals along with their response vectors, $\hat{\theta}$, and fit information. We can examine this information to identify the reason for misfit. A snippet of BIGSTEPS's TABLE 7.1 is shown in Table 3.7. The table's format contains a column labeled NUMBER that refers to the case's ordinal position in the file, followed by the person's label (if any), location estimate (MEASURE), and INFIT ZSTD and OUTFIT ZSTD values. The subsequent line beginning with RESPONSE contains the person's item responses, and the responses' corresponding residuals are on the line labeled Z-RESIDUAL. For example, the first person listed is person 358 with a $\hat{\theta} = 1.95$, INFIT ZSTD = 1.5, OUTFIT ZSTD = 9.8, and a response pattern of 01111.

From Table 3.7 we see that the OUTFIT ZSTD value of 9.8 is associated with persons who, although they have high proficiency estimates ($\hat{\theta} = 1.95$), incorrectly answered

TABLE 3.7. Abridged Person Response Vector Misfit Information for Poorly Fitting People

TABLE 7.1 rasch calibration math data						
INPUT: 19601 CASES, 5 ITEMS ANALYZED: 15525 CASES, 5 ITEMS, 10 CATS v2.82						

TABLE OF POORLY FITTING CASES (ITEMS IN ENTRY ORDER)						
NUMBER - NAME -- POSITION -----	MEASURE	- INFIT (ZSTD)	OUTFIT			
358 358		1.95	1.5	A	9.8	
RESPONSE: 1: 0 1 1 1 1						
Z-RESIDUAL: -9						
1236 1,236		1.95	1.5	B	9.8	
RESPONSE: 1: 0 1 1 1 1						
Z-RESIDUAL: -9						
2022 2,022		1.95	1.5	C	9.8	
RESPONSE: 1: 0 1 1 1 1						
Z-RESIDUAL: -9						
2349 2,349		1.95	1.5	D	9.8	
RESPONSE: 1: 0 1 1 1 1						
Z-RESIDUAL: -9						
2800 2,800		1.95	1.5	E	9.8	
RESPONSE: 1: 0 1 1 1 1						
Z-RESIDUAL: -9						
:						
<more examinees>						
:						

the easiest item ($\hat{\delta}_1 = -2.21$) and correctly answered the remaining more difficult items; Table 2.1 shows there are 40 individuals with this response vector (i.e., 01111; $X_i = 4$). The incorrect response to the easiest item may have been due to carelessness, inattentiveness, or another factor. The Z-RESIDUAL's negative value for item 1 indicates that, according to the model, these persons were expected to respond correctly but did not; the magnitude of Z-RESIDUAL indicates the degree of discrepancy.

What should we do with these 40 misfitting individuals? If we are concerned that these persons are adversely affecting the item location estimation, then these individuals could be removed, the instrument recalibrated, and these persons assigned the recalibrated $\hat{\theta}$ corresponding to an observed score of 4. However, these 40 examinees represent 0.2% of the sample, and so their removal's potential impact on the accuracy of item parameter estimation should be negligible. This should not be interpreted to mean that the item parameter estimates themselves would not change. Most likely the removal of these individuals would affect the $\hat{\delta}_j$ s and, as such, would also affect the fit statistics for all the items, not only those for item 1.¹¹ However, the INFIT ZSTD for these 40 persons is 1.5, which indicates less reason to be concerned about these individuals. In other words, these persons are responding relatively consistently to items near their estimated ability. Given this rationale and that the person OUTFIT value's magnitude is driven by an anomalous response to just one item, these examinees are retained.

Invariance Assessment

Theoretically, IRT item parameters are invariant. In this regard, *invariance* refers to one or more parameter metrics that are interchangeable within a permissible transformation (cf. Rupp & Zumbo, 2006). However, whether invariance is realized in practice (i.e., with parameter estimates) is contingent on the degree of model–data fit. Therefore, the presence of invariance can be used as part of a model–data fit investigation. The quality of model–data fit may be assessed by randomly dividing the calibration sample into two subsamples.¹² Each subsample is separately calibrated, and their item parameter estimates are compared to determine their degree of linearity. One measure of agreement between the two samples' estimates is the Pearson product-moment correlation coefficient.

The first step in using invariance for assessing model–data fit is to randomly divide the sample in half. One way to do this is to assign a uniform random number to each individual. To create two subsamples of approximately equal size, this random number is compared to 0.50. If the random number for a person is greater than 0.50, then the individual belongs to one subsample; otherwise the person belongs to the other subsample. Applying this process to the example data produced two subsamples, one with 9780 persons and the other with 9821 individuals. Independent calibrations of each subsample resulted in $\hat{\delta}_1 = -2.18$, $\hat{\delta}_2 = -0.31$, $\hat{\delta}_3 = 0.22$, $\hat{\delta}_4 = 1.02$, and $\hat{\delta}_5 = 1.26$ for subsample 1 and $\hat{\delta}_1 = -2.25$, $\hat{\delta}_2 = -0.22$, $\hat{\delta}_3 = 0.18$, $\hat{\delta}_4 = 1.02$, and $\hat{\delta}_5 = 1.27$ for subsample 2. Therefore, subsample 1's item location estimates are not equal to the corresponding estimates in subsample 2, although they show the same rank order. The Pearson correlation coefficient between these two sets of item parameter estimates is

0.9991. Therefore, we have evidence of invariance across these subsamples. This result provides additional evidence of model–data fit. (It should be noted that a large Pearson correlation coefficient is a necessary, but not a sufficient, condition for invariance; e.g., see Rupp & Zumbo [2004].) For completeness, the Pearson correlation coefficient between the first subsample’s estimates and those of the “full” sample ($N = 15,525$) is 0.9998, and for the second subsample’s $\hat{\theta}$ s the correlation with the full-sample $\hat{\theta}$ s is 0.9998. The implication of the magnitude of these correlations is that we can use a linear transformation to convert the estimates on one metric to that of the other metric without any loss of information concerning model–data fit or person and item location estimates.

Why do the estimates across subsamples not agree with one another in magnitude even though there is apparent model–data fit? The primary reason is that, as stated above, the latent variable continuum’s metric is not absolute, but rather is defined with respect to the information available for estimation. Each sample defines the continuum’s metric for the item location estimates. A second reason is that the estimates have some error as indicated by $s_e(\hat{\theta})$. The relative nature of the metrics is not problematic as long as the metrics are highly linearly related. When this is true, then we can use a linear transformation to convert from one metric to the other.

How do the “by-hand” calculations from Chapter 2 (Table 2.3) compare to those of BIGSTEPS? A comparison of full-sample BIGSTEPS’s $\hat{\theta}$ s with those obtained by hand reveals that the values do not correspond exactly. For example, the by-hand estimated item 1 location is -1.90 , but BIGSTEPS estimated its location to be -2.18 . However, the two sets of item parameter estimates are highly linearly related ($r = 0.9998$). The primary reason for the difference between our by-hand estimates and BIGSTEPS’s is due to algorithmic differences. In essence, our hand calculations performed only step 1 of the JMLE process and did not “ping-pong” between JMLE’s steps 1 and 2. In addition, we used a stricter convergence criterion of 0.0001 in our hand-based calculations than did BIGSTEPS (i.e., $LCONV = 0.01$).

Of course, if the metric for the hand-based calculations does not match that of BIGSTEPS, then there is no reason to believe the hand-based $\hat{\theta}$ s will be equal to BIGSTEPS’s $\hat{\theta}$ s (i.e., the $\hat{\theta}$ s are on the same metric as the $\hat{\theta}$ s). However, if the two metrics are highly linearly related, then so will the corresponding $\hat{\theta}$ s. A comparison of the $\hat{\theta}$ s from Table 2.2 and those from BIGSTEPS (i.e., for $X = 1, 2, 3, 4$ the $\hat{\theta}$ s are $-2.00, -0.43, 0.73$, and 1.95 , respectively) shows the two sets of estimates do not match. However, the two sets of $\hat{\theta}$ s are highly linearly related ($r = 0.9990$).

Example: Application of the Rasch Model to the Mathematics Data, JMLE, mixRasch

mixRasch is an R package that fits the polytomous mixture (Rasch) model (see Chapter 1 and Appendix F, “Mixture Models”) using JMLE. The dichotomous Rasch model is a special case of a polytomous mixture model when we have one latent class

and binary response data. Thus, we can use `mixRasch` to fit our Rasch model. Table 3.8 shows our R session; our data have already been read into the workspace as shown in Table 3.1.

We invoke the `mixRasch` function by `mixRasch(mathdata, steps=1, info.fit=T)` and with the output object `rasch`. By default, `mixRasch` fits a one-latent class rating scale model (see Chapter 7 for information on the rating scale model). Thus, our arguments to `mixRasch` are the data frame (`mathdata`), the number of thresholds (which for binary data is 1; `steps=1`), and that we would like the information criteria to be calculated (`info.fit=T`; we use this as a way to obtain the number of examinees used in estimation). As can be seen, `mixRasch` performed eight iterations before achieving convergence. By using `rasch$converge.flag`, we know we have a converged solution because the variable is set true; alternatively, we can pass the output object to the `getEstDetails` function (e.g., `getEstDetails(rasch)`). As mentioned above, with JMLE all cases with zero variance response vectors (i.e., perfect scores, zero scores) are removed from estimation. Therefore, to determine the number of cases used for estimation, we access this information in `info.fit` (`rasch$info.fit`); alternatively, we can directly access this information by `rasch$info.fit$N.persons`. The `$N.persons` variable shows that 15,525 examinees were used for estimation, as was the case with BIGSTEPS. Model-level fit information is also provided in terms of the information indices AIC, BIC, and CAIC. These indices provide relative fit information for model comparison by taking into account model complexity among other things; these indices are based on the $\ln L$ value of $-34,844.18$. Because we do not have any model(s) for a comparison model, we ignore the AIC, BIC, and CAIC values (we discuss these indices in Chapter 5).

We can obtain our item location estimates by simply typing the output object name, `rasch`, or we can use the `getItemDetails` function (e.g., for item 1: `getItemDetails(rasch, "I1." class=1)`). (Directly accessing the `item.par` variable (i.e., `rasch$item.par`) will provide the item location estimates with corresponding standard errors, item-level fit information, p -values, and point-biserial and biserial correlations.) As can be seen, our item location estimates are $\hat{\delta}_1 = -2.767$, $\hat{\delta}_2 = -0.327$, $\hat{\delta}_3 = 0.245$, $\hat{\delta}_4 = 1.266$, and $\hat{\delta}_5 = 1.583$, with standard errors ranging from 0.019 to 0.029. A comparison of these estimates to those of BIGSTEPS (cf. Table 3.4) shows a perfect correlation. Although the `mixRasch` item estimates are close to those of BIGSTEPS they do not match exactly. This is because the `mixRasch` estimates are on a slightly different metric than those of BIGSTEPS. Specifically, the mean and standard deviation for `mixRasch` are 0.0000 and 1.7271, respectively, whereas for BIGSTEPS the `M` is 0.0020 and the `SD` is 1.3806; Table 3.4's S.D. of 1.24 is a noninferential `SD`. When we linearly transform the `mixRasch` metric to that of BIGSTEPS, the estimates match exactly (except for rounding error).

Also shown in our item parameter estimates table is fit information. This fit information may also be presented graphically by using the `itemFitPlot(. . .)` function (Figure 3.1). The shaded portions indicate INFIT or OUTFIT values outside of the 0.7–1.3 range. Because these INFIT and OUTFIT values correspond to those of BIGSTEPS, the interpretation given above applies here. For example, using the common

TABLE 3.8. Rasch Session for the Rasch Calibration of the Mathematics Data

```

> # This is a continuation of the session from Table 3.1

> library(mixRasch)
> packageVersion("mixRasch")
[1] '1.1'

> rasch=mixRasch(mathdata,steps=1, info.fit=T)
  Iteration: 1, Largest Parameter Change: 1.579118
  Iteration: 2, Largest Parameter Change: 0.7859293
  Iteration: 3, Largest Parameter Change: 0.2850698
  Iteration: 4, Largest Parameter Change: 0.08597095
  Iteration: 5, Largest Parameter Change: 0.02286737
  Iteration: 6, Largest Parameter Change: 0.005810631
  Iteration: 7, Largest Parameter Change: 0.001458762
  Iteration: 8, Largest Parameter Change: 0.0003652269

> rasch$converge.flag
[1] TRUE

> rasch$info.fit
$AIC
[1] 69704.36

$BIC
[1] 69765.56

$CAIC
[1] 69773.56

$loglik
[1] -34844.18

$N parms
[1] 8

$N persons
[1] 15525 # this matches BIGSTEPS

> rasch # one way to get item estimates
   difficulty      se infit     in.Z outfit    out.Z
i01     -2.767 0.029 1.069    3.674   1.173    3.269
i02     -0.327 0.020 0.857   -15.909   0.784   -16.593
i03      0.245 0.019 0.934   -7.582   0.909   -7.211
i04      1.266 0.020 0.979   -2.317   1.047    2.374
i05      1.583 0.021 1.174   17.337   1.391   14.848

Time difference of 4.445624 secs

Number of iterations: 8

> itemFitPlot(rasch, fitStat="infit", colTheme="greys") # See Figure 3.1
> itemFitPlot(rasch, fitStat="outfit", colTheme="greys") # See Figure 3.1

```

(continued)

TABLE 3.8. (continued)

```

> # See Figure 3.2 for the following plots
> rICC(rasch$item.par$delta[,1], rasch$person.par$theta,mathdata[,1], empICC=TRUE)
> rICC(rasch$item.par$delta[,2], rasch$person.par$theta,mathdata[,2], empICC=TRUE)
> rICC(rasch$item.par$delta[,3], rasch$person.par$theta,mathdata[,3], empICC=TRUE)
> rICC(rasch$item.par$delta[,4], rasch$person.par$theta,mathdata[,4], empICC=TRUE)
> rICC(rasch$item.par$delta[,5], rasch$person.par$theta,mathdata[,5], empICC=TRUE)

> rasch$person.par
   theta SE.theta   r     infit      in.Z    outfit      out.Z
 1 -0.4254419 1.125744 2 0.5520816 -38.880309 0.4059338 -37.4966929
 2  0.7303919 1.055163 3 0.4878208 -64.390548 0.4009607 -27.8697135
 3 -2.0008133 1.417238 1 0.2750098 -55.534534 0.1648530 -38.6602541
 4  0.7303919 1.055163 3 0.4878208 -64.390548 0.4009607 -27.8697135
 5  0.7303919 1.055163 3 1.3183112 13.702108 1.1318614  2.1189458
 6  0.7303919 1.055163 3 0.4878208 -64.390548 0.4009607 -27.8697135
 :
 
> peopleRasch =matrix(rasch$person.par[,1])
> head(peopleRasch)
   [,1]
[1,] -0.4254419
[2,]  0.7303919
[3,] -2.0008133
[4,]  0.7303919
[5,]  0.7303919
[6,]  0.7303919

> tail(peopleRasch,4)
   [,1]
[19598,] 1.950402
[19599,] 3.444043
[19600,] 1.950402
[19601,] 1.950402

> # for comparison with Table 3.6
> peopleRasch[9]    # X = 5, perfect score person
[1] 3.444043

> # X = 4
> peopleRasch[7]    # X = 4
[1] 1.950402

> # X = 3
> peopleRasch[2]    # X = 3
[1] 0.7303919

> # X = 2
> peopleRasch[1]    # X = 2
[1] -0.4254419

```

(continued)

TABLE 3.8. (continued)

```

> # X = 1
> peopleRasch[3]    # X = 1
[1] -2.000813

> peopleRasch[19414]   # X = 0, zero score person
[1] -3.884187

> peopleRasch[19585]   # X = 0
[1] -3.884187

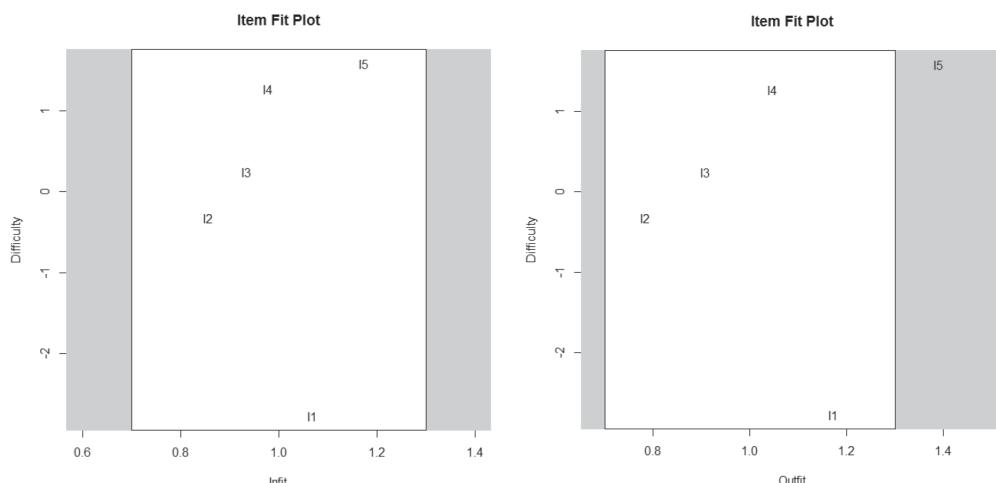
> # See Figure 3.3 for the following plot

> personItemPlot(rasch, nBreaks=18, plotTitle="Person Item Histogram", xlab =
  "Relative Frequency", ylab = "Ability", col = c("darkgrey", "lightgrey"),
  makeLegend=TRUE, legendLabels=c("items", "people"), legendLoc="bottomright")

```

cutoff points of 0.7 and 1.3, there does not appear to be any apparent serious fit problems because the fit statistics are within acceptable bounds; refer to the BIGSTEPS analysis for comments concerning item 5. (Please recall our comments about common cutoffs in our discussion of Table 3.4.)

`mixRasch` allows the plotting of the predicted IRFs with or without the empirical IRF superimposed using the `rICC(. . .)` function, where the first and second arguments are the item location and person location estimates, respectively, and the third and fourth optional arguments indicate if the empirical IRF is desired (`empICC=TRUE`) and the item of interest (e.g., for item 1: `mathdata[,1]`). The empirical IRFs use the $\hat{\theta}$ s for plotting. As such, with only five items there are only four unique person location estimates, and the empirical IRFs are somewhat crude and of limited value; `mixRasch`

**FIGURE 3.1.** Item INFIT (left) and OUTFIT (right) plots.

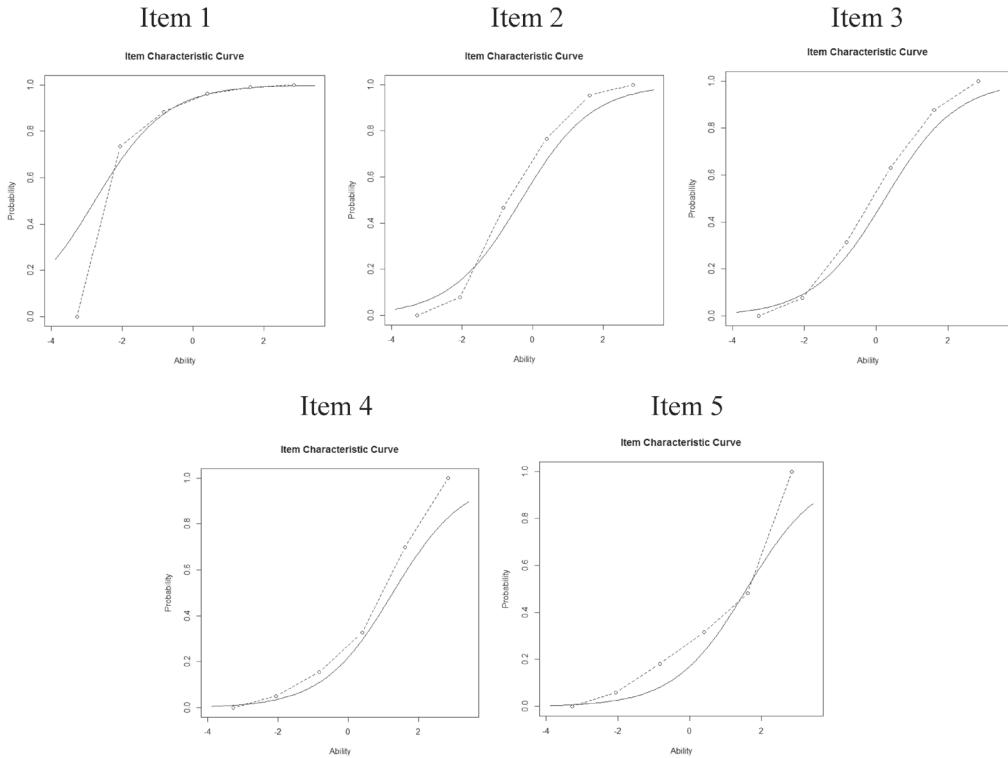


FIGURE 3.2. Empirical and predicted IRFs.

uses its nonestimated $\hat{\theta}$ s for the $X = 0$ and $X = 5$. Nevertheless, for pedagogical reasons, we present these plots in Figure 3.2. For example, for item 1 the empirical IRF (dash line) is similar to the predicted IRF (solid line) from roughly -2 to 2 . The leftmost and rightmost empirical points (circles) represent the pseudo $\hat{\theta}$ s for the $X = 0$ and $X = 5$, respectively. Because these particular $\hat{\theta}$ s are somewhat arbitrary, they should not be given much weight.

To obtain our examinee location estimates, we use the variable name `person.par` with the output object (`rasch$person.par`). This provides us with our $\hat{\theta}$ s, $s_e(\hat{\theta})$ s, X es, and fit information. For example, for our first examinee $\hat{\theta}_1 = -0.4254419$, $s_e(\hat{\theta}_1) = 1.125744$, $X_1 = r = 2$, and unstandardized (INFIT = 0.5520816, OUTFIT = 0.4059338) as well as standardized fit information (in.Z = -38.880309, out.Z = -37.4966929). Except for the standardized fit statistics, these match those of BIGSTEPS (cf. Table 3.6 where examinee 1 is midway down the table). For convenience we reformat our examinee location estimates using the `matrix` function (`matrix(rasch$person.par[,1])`) and display the first six (`head`) and last (`tail`) four examinees. We also display the $\hat{\theta}$ s for the examinees shown in Table 3.6. Therefore, we know that for the observed scores of 5, 4, 3, 2, 1, and 0, the corresponding person estimates are 3.444043, 1.950402, 0.7303919, -0.4254419, -2.000813, and -3.884187, respectively. For the nonzero variance

response vectors, these values correspond to those of BIGSTEPS's (cf. Table 3.6). The M and SD for the 15,525 unique $\hat{\theta}$ s are 0.1849 and 1.3951, respectively; for BIGSTEPS $M = 0.1836$ and $SD = 1.3953$.

Similar to BIGSTEPS, mixRasch provides $\hat{\theta}$ s for the zero variance response vectors ($X = 0, X = 5$) by using a kludge. Specifically, it subtracts a fractional score point from the perfect score and adds the same fractional score point to the 0 score to create scores that are estimable. By default, this fractional score point is 0.3 (the control parameter `treat.extreme` can be used to change this value). Therefore, a perfect score of 5 is converted to 4.7 and a zero score becomes 0.3. Hence, the $\hat{\theta}$ s of -3.884187 and 3.444043 are *not* the MLE for a zero score and a perfect score, respectively. Rather, these are artifacts of the particular fractional score point value used. If we set `treat.extreme = 0.5`, then we obtain the same pseudo $\hat{\theta}$ s as BIGSTEPS and obtain $\hat{\theta}$ s of -3.192352 and 2.855807 for $X = 0$ and $X = 5$, respectively. The choice of the fractional value does not affect the item location estimates.

An advantage of the Rasch model that is useful in scale development is that items and people are located on the same continuum. As a result, it is possible to ascertain how well targeted an instrument is for a sample by using a Variable Map (Engelhard, 2013; Smith, 2004); a Variable Map is also called an Item-Person Map, Person-Item Map, Joint Distribution Map, and Wright Map (cf. Linacre, 2001a; Linacre & Wright, 2001; Mair et al., 2018; Stone, Wright, & Stenner, 1999; Torres Irribarra & Freund, 2014). The Variable Map shows the person and item distributions on the theta scale. Typically, this graphic consists of a horizontal histogram for persons and one for items side-by-side, with the ordinate reflecting the latent variable and the abscissa the frequency or some transformation thereof. The top of the graph represents "more of" and the bottom represents "less of" whatever the latent variable is measuring. With respect to our math test, the top would reflect greater math ability for people and difficult math items, whereas the bottom would indicate lesser math ability for people and easier math items. If relevant, one can impose one or more standards (e.g., for passing, certification) and/or item content to get a sense of how the instrument is functioning and/or where specific content is tapping the continuum.

By using the `personItemPlot` function, we can create a variable map (Figure 3.3). As can be seen, the item location estimate distribution (left side) maps well to the person location estimate distribution (right side). (Because the location of the bottom ($X = 0$) and top ($X = 5$) person bars are based on the kludge values discussed above, we do not give these two bars as much weight as the others.) The sum of the proportions of the person bars above an item's location gives us an idea of how many people would answer the item correctly. When an item and people are at the same point on the ordinate (e.g., the second item from the bottom), then one expects the probability of correct response on this item to be in the neighborhood of 0.5 for the people located in the corresponding person bar (third bar from the bottom). From a test design perspective, we see that our examination would benefit from items targeted to fall within the -2.5 to -1.0 gap. The benefits of using a variable map increase as the instrument lengthens and the sample size increases. (In Chapter 7 we introduce BIGSTEPS's variable map.)

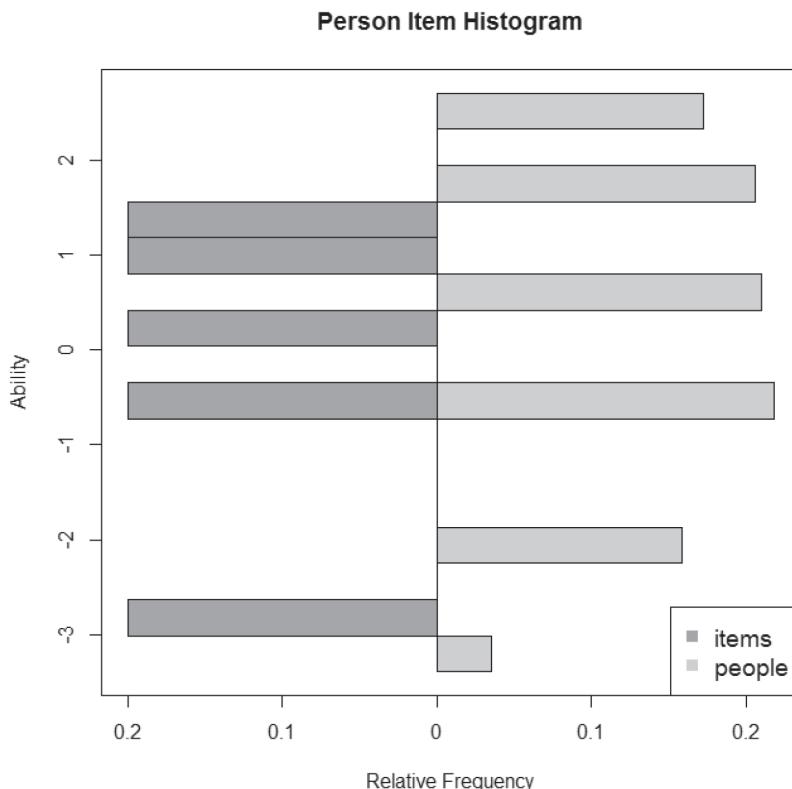


FIGURE 3.3. Variable map.

Validity Evidence

Recall that prior to estimating the item and person locations, we assumed that we had content validity evidence for our instrument (see Bandalos, 2018). Thus, we subjected our instrument to a rational analysis to determine whether the items were measuring mathematics proficiency. Moreover, we examined the instrument to make sure there were no items that, owing to nonmathematical proficiency factor(s), could disadvantage classes of individuals. Although we have Rasch model–data fit, it is still necessary to acquire additional validity evidence for the $\hat{\theta}$ s in terms of the mathematics construct as well as criterion-related validity.

From a traditional perspective, we may seek to obtain construct- and criterion-related validity evidence. For instance, if there were additional measures of mathematics proficiency that could serve as criteria (e.g., performance in mathematics courses), then one could obtain criterion-related validity evidence by regressing each of these measures on our instrument's $\hat{\theta}$ s. Obtaining construct validity evidence can involve making predictions about differential performance between groups and determining whether the instrument correctly differentiates between the groups. For example, if the instrument purports to measure mathematics at and beyond algebra, then it should

be able to identify individuals who have a knowledge of algebra and higher-level mathematics from those who do not. Additional approaches may include the application of the multitrait–multimethod validity paradigm (Campbell & Fiske, 1959).

As part of this validation process, we should also perform a *differential item functioning* (DIF) analysis (Chapter 12 contains an overview of DIF). DIF analysis determines whether performance on any of the items differs for certain groups of individuals (e.g., females vs. males) after controlling for differences in person location (e.g., math ability). Although this is discussed in detail in Chapter 12, the premise of DIF is that if one matches, for example, males and females on mathematics proficiency, then the probability of correctly responding to an item should be the same for females and males. However, if we find that the probability of correctly responding to an item for males is less than that for females (or vice versa), then the item is functioning differentially across gender. In this case, the item would appear to be measuring not only the construct of interest, mathematics proficiency, but also an additional tangential or nuisance factor(s) associated with gender (see Ackerman, 1992; Kok, 1988).

Summary of the Application of the Rasch Model

In general, the application of the Rasch model proceeds by administering a set of items to a sample of individuals, performing a model–data fit analysis and examining items and persons for poor model–data fit. The fit analysis uses both graphical and statistical fit indices. The examination of items flagged as poorly fitting is concerned with discovering why an item is identified as a poor fit to the model. For instance, if the item misfit is due to its unique discrimination capacity, then this may reflect an item that is being interpreted in more than one way by the respondents. Rewording the item could potentially eliminate this ambiguity. After this examination one may conclude that such poorly fitting items should be removed from the instrument. However, simply because an item is identified as exhibiting poor fit does not automatically result in its removal. It is the examination of the poorly fitting items that should lead to an item's retention or removal. Also, because a model represents an approximation, a certain degree of misfit is expected and tolerated.

The argument for item removal relies on the observation that the item is not functioning in a fashion consistent with the model. Some have used the removal of items as a criticism of the Rasch model. However, this strategy is not different from that used in other routinely used techniques for instrument development. For example, if one performs a factor analysis and finds an item that has a low loading on the instrument's factors, then rather than introducing a single-item factor, the item is typically discarded. Similarly, in traditional item analysis items may be removed because of a valid negative biserial (i.e., the sign is not due to an incorrect keying of the correct response). In this situation, the item is not behaving in a fashion consistent with conventional wisdom. Furthermore, one approach to scale development relies on discarding items to maximize the observed scores' reliability. Therefore, the removal of items to improve an

instrument's psychometric properties is not without precedent and is acceptable with other psychometric analyses (cf. Samejima, 1973a).

After the Rasch model–data fit process is completed, one has constructed a scale for measuring a variable. This scale fits the Rasch model to some degree in the same sense that Guttman (1950) accepted some degree of deviation from a perfect Guttman scale in his Scalogram method and Walker (1931) accepted deviation from a unig test. In other words, the Rasch approach to measurement is found in other measurement procedures.

To summarize our analyses, the nonlinear factor analysis provided support that a unidimensional model of the data is a reasonable representation of the data. The INFIT and OUTFIT fit statistics at the individual item and person level, the subsequent response vector analysis to investigate large OUTFIT ZSTD values, and the invariance assessment all provide evidence supporting model–data fit. Presumably, validity evidence would also support the application of the Rasch model to our mathematics data. If so, then, these data would have been successfully calibrated using the Rasch model. The calibration indicates that the instrument does and will perform comparatively better in estimating persons toward the middle and lower end of the mathematics continuum than those toward its upper end.

Summary

Estimating item and person locations may be accomplished by various methods. One approach is to “simultaneously” estimate both item and person locations by maximizing the joint likelihood function of persons and items. This method is called joint maximum likelihood estimation (JMLE). The procedure begins with provisional estimates of person locations, and these are treated as “known” for estimating the items’ parameters via Newton’s method. Once convergence is obtained for the item parameter estimates, these estimated item parameters are treated as “known” and the person locations are estimated via Newton’s method. Upon achieving convergence for the persons, these person locations are presumably more accurate than the provisional person location estimates initially used for item parameter estimation. Therefore, the improved person location estimates are treated as “known” and the item parameters are re-estimated. Similarly, the improved item parameter estimates are used to re-estimate the person parameters. The JMEL method ping-pongs between the two stages until the difference between successive iterations is sufficiently small to satisfy the convergence criteria.

One implication of not knowing either the item or the person parameters is that the continuum's metric is indeterminate. If we knew either set of parameters (i.e., persons or items), then the metric would be determined by the parameters' values. However, because we do not know either set of parameters, the metric needs to be fixed in some fashion to identify the model. Two common approaches are person centering and item centering. With both approaches, sample information is used for fixing the metric when estimating person and/or item parameters. Therefore, when different samples are used to estimate an instrument's item parameters, the resulting estimates may not necessarily

be identical. However, if there is model–data fit, then the estimates are linearly related and one metric may be transformed to the other.

Theoretically, the person and item parameters are invariant, but we have only estimates of these parameters. However, if we have model–data fit, then we expect to observe invariance of our parameter estimates. Thus, part of our model–data fit analysis involves examining the invariance of our parameter estimates.

The application of an IRT model involves an analysis of model–data fit. All IRT models make dimensionality assumptions. Therefore, one of the first steps in our fit analysis is to determine whether, for example, a unidimensional model is an accurate representation of the data. Although various approaches can be used for dimensionality assessment, with our binary response data we used a nonlinear factor analysis to mitigate the possibility of obtaining “difficulty” factors. (We consider obtaining validity evidence as part of dimensionality assessment.) Our second model–data fit step uses item/person/model fit measures. For instance, our calibration program’s INFIT and OUTFIT indices are used to determine if the items and people were behaving consistently with the model. The third step in our model–data fit analysis looks at the degree of linearity between the item parameter estimates from the separate calibrations of subsamples to examine invariance of our parameter estimates. For example, if we randomly create two subsamples and calibrate each sample, then a large correlation between the item location estimates from the two subsamples would provide evidence of the invariance of our estimates. If our estimates are highly linearly related, then not only would we have evidence supporting the invariance of the parameter estimates, but also by implication, model–data fit.

In the next chapter, we introduce a different approach for performing IRT parameter estimation, marginal maximum likelihood estimation (MMLE). This approach is used in estimation programs such as BILOG-3 (Mislevy & Bock, 1997), BILOG-MG, flexMIRT, PARSCALE, ConQuest (Wu, Adams, & Wilson, 1997), and OPLM, to name a few. Unlike JMLE, MMLE separates item parameter estimation from person parameter estimation. Parallel to this chapter’s structure, we first discuss the MMLE estimation approach, followed by its application to the mathematics data using BILOG-MG and the R package mirt. As part of this example, we introduce the concepts of the metric alignment and an instrument’s total characteristic curve.

Notes

1. An alternative is the *standard item* approach. Using this strategy, we select a convenient item (e.g., the first item) and set its value to some fixed value, say 0.0. The remaining items, estimated locations, are expressed as multiples of the standard item. For example, in a proficiency situation the standard item reflects a unit of difficulty and all other item location estimates reflect multiples of this degree of difficulty, with some items more difficult and others less difficult than the standard item (cf. Rasch, 1966). This approach fixes the metric around the standard item and thereby resolves the metric indeterminacy.

2. This recommendation is based on using JMLE. Moreover, Lord (1983a), using LOGIST, looked at the bias in estimating $\hat{\theta}$ using either the 1PL model or a more complicated model, the two-parameter model (see Chapter 5). He concluded that for 10- to 15-item instruments the unweighted observed score (i.e., $X_i = \sum x_{ij}$) used in the 1PL model was slightly superior to the weighted observed score used in the two-parameter model when the item parameters were estimated on the basis of “less than 100 or 200” cases (Lord, 1983a, p. 58).
3. NOHARM4 requires the creation of an ASCII (text) input file. The command input files for performing one- and two-dimensional analyses for NOHARM4 are shown in Table 3.9; NOHARM4 can also be accessed using sirt’s R2noharm function. The two command files have the same structure and differ only in the second program parameter on the second line. We have on the first command line a title and on line 2 the program parameters (e.g., number of items, number of dimensions, number of cases, case (0) or product-moment matrix (1) entry). For instance, “5 1 19601 0 1 0 0 0” specifies there are five items; to perform a one-dimensional analysis, there are 19,601 cases; the input is case data; an exploratory analysis (rather than confirmatory analysis) is requested; NOHARM4 should generate its starting values; and the correlation, covariance, and residual matrices should be printed, respectively. On with line 3, we specify the value to use for the IRF’s lower

TABLE 3.9. One- and Two-Dimensional Input Command File; Product-Moment Input

One-dimensional input

```
EXPLORATORY ANALYSIS, math.dat (case data input), 1D
 5 1 19601 0 1 0 0 0
 0 0 0 0 0
 1 1 0 0 0
 1 1 1 0 0
 1 0 0 0 0
 1 1 1 0 0
 1 0 1 1 0
 1 1 1 0 0
 :
 1 1 1 1 0
```

Two-dimensional input

```
EXPLORATORY ANALYSIS, math.dat (case data input), 2D
 5 2 19601 0 1 0 0 0
 0 0 0 0 0
 1 1 0 0 0
 1 1 1 0 0
 1 0 0 0 0
 1 1 1 0 0
 1 0 1 1 0
 1 1 1 0 0
 :
 1 1 1 1 0
```

asymptote value for each of the five items. In this case, the lower asymptote for each item is set to 0 because we are assuming the response data are not influenced by guessing. Following these lines are our case data. After creating the command file, we start NOHARM4. A dialog (Figure 3.4) allows us to specify the command file (Step 1), a filename for the output file (Step 2), and estimation options (Step 3), and to execute the command file (Step 4). Additionally, the program's manual is accessible by clicking on the User's Guide button.

Tables 3.10 and 3.11 contain the corresponding abridged output; Appendix G, "Example: NOHARM Unidimensional Calibration," contains additional output. These results match those of *noharm.sirt*, except that the residual matrix is opposite in sign; RMS is labeled RMSR in *noharm.sirt*. The beginning of the output contains echoes of the input specifications. Beginning with "Residual Matrix (lower off-diagonals)," we find NOHARM4's fit information.

4. Analogous to principal component's or principal axis's analysis of a correlation matrix, NOHARM analyzes the raw product-moment matrix ($\underline{\mathbf{P}}$). Rather than have NOHARM calculate $\underline{\mathbf{P}}$ from case data, one can calculate a data set's product-moment matrix once and provide it to NOHARM for repeated analyses (e.g., for examination of multiple dimensional solutions). The raw product-moment matrix is obtained by $\underline{\mathbf{X}}'\underline{\mathbf{X}}(1/N)$, where N is the number of cases and $\underline{\mathbf{X}}$ is the data matrix. If we let $\underline{\mathbf{X}}$ consist of binary responses, then $\underline{\mathbf{P}}$ contains the item means or traditional item difficulties, P_j s, along its main diagonal and the sums of raw product terms divided by N as its off-diagonal elements. For example, for a three-item instrument, the raw product-moment matrix is

$$\underline{\mathbf{P}} = \begin{bmatrix} P_1 & (\sum X_1 X_2)/N & (\sum X_1 X_3)/N \\ (\sum X_2 X_1)/N & P_2 & (\sum X_2 X_3)/N \\ (\sum X_3 X_1)/N & (\sum X_3 X_2)/N & P_3 \end{bmatrix}, \quad (3.4)$$

where $P_1 = \sum X_1 / N$, $P_2 = \sum X_2 / N$, $P_3 = \sum X_3 / N$, and all subscripts refer to items. As an example of using $\underline{\mathbf{P}}$ for our analyses, we use R to obtain $\underline{\mathbf{P}}$. The R session is

```
> mathdata=read.table("math.dat",header=F)
> x=data.matrix(mathdata)           # convert data frame to numeric matrix
> xtranspose=t(x)                 # transpose of x
> XtransposeX=xtranspose%*%x     # calculate  $\underline{\mathbf{X}}'\underline{\mathbf{X}}$ 

> pmm =XtransposeX*(1/19601)      # pmm =  $\underline{\mathbf{X}}'\underline{\mathbf{X}}(1/N)$ 

> pmm
    i01        i02        i03        i04        i05
i01 0.8874547 0.6068058 0.5307382 0.4013571 0.3604918
i02 0.6068058 0.6440488 0.4423244 0.3522779 0.3024846
i03 0.5307382 0.4423244 0.5659915 0.3173307 0.2753431
i04 0.4013571 0.3522779 0.3173307 0.4269680 0.2225397
i05 0.3604918 0.3024846 0.2753431 0.2225397 0.3873272
>
```

TABLE 3.10. Abridged One-Dimensional Analysis Output

N O H A R M					
Fitting a (multidimensional) Normal Ogive by Harmonic Analysis - Robust Method					
Input File : math1Dpm.cmd					
Title : EXPLORATORY ANALYSIS, math.dat (product moment input), 1D					
Number of items = 5					
Number of dimensions = 1					
Number of subjects = 19601					
An exploratory solution has been requested.					
Sample Product-Moment Matrix					
1 2 3 4 5					
1	0.887				
2	0.607	0.531			
3	0.401	0.360	0.607		
4	0.644	0.442	0.352	0.302	
5	0.531	0.442	0.566	0.317	0.275
Item Covariance Matrix					
1 2 3 4 5					
1	0.100				
2	0.035	0.229			
3	0.028	0.078	0.246		
4	0.022	0.077	0.076	0.245	
5	0.017	0.053	0.056	0.057	0.237
Fixed Guesses					
1 2 3 4 5					
0.0	0.0	0.0	0.0	0.0	
:					=====
Results					
=====					
Success. The job converged to the specified criterion.					
Final Constants					
1 2 3 4 5					
1.438	0.551	0.229	-0.256	-0.335	
:					
Residual Matrix (lower off-diagonals)					
1 2 3 4					
2	-0.005				
3	7.9e-5	0.001			
4	0.004	-5.0e-4	1.5e-4		
5	0.003	0.002	-0.002	-0.001	
Sum of squares of residuals (lower off-diagonals) = 0.0000543					
Root mean square of residuals (lower off-diagonals) = 0.0023298 ← The RMSR					
Tanaka index of goodness of fit = 0.9996365 ← The GFI					

TABLE 3.11. Abridged Two-Dimensional Analysis Output

N O H A R M

Fitting a (multidimensional) Normal Ogive
by Harmonic Analysis - Robust Method

Input File : math2Dpm.cmd

Title : EXPLORATORY ANALYSIS, math.dat (product moment input), 2D

Number of items = 5

Number of dimensions = 2

Number of subjects = 19601

An exploratory solution has been requested.

Sample Product-Moment Matrix

	1	2	3	4	5
1	0.887				
2	0.607	0.531			
3	0.401	0.360	0.607		
4	0.644	0.442	0.352	0.302	
5	0.531	0.442	0.566	0.317	0.275

Item Covariance Matrix

	1	2	3	4	5
1	0.100				
2	0.136	0.249			
3	-0.137	0.038	0.239		
4	0.376	0.282	0.169	0.211	
5	0.286	0.296	0.399	0.234	0.200

Fixed Guesses

1	2	3	4	5
0.0	0.0	0.0	0.0	0.0
:				

=====

Results

=====

Success. The job converged to the specified criterion.

:

Residual Matrix (lower off-diagonals)

	1	2	3	4
2	-4.3e-5			
3	-7.0e-5	1.8e-4		
4	2.6e-4	-0.001	0.001	
5	-2.2e-4	0.001	-0.001	2.1e-4

Sum of squares of residuals (lower off-diagonals) = **5.98e-006**

Root mean square of residuals (lower off-diagonals) = 0.0007732 ← The RMSR

Tanaka index of goodness of fit = 0.9999600 ← The GFI

:



FIGURE 3.4. NOHARM4 input dialog.

For our NOHARM analysis using `sirt`, we specify the product-moment matrix ($\text{pm}=\text{pmm}$) along with the number of cases ($N=19601$); please see the text for a description of the remaining arguments. The function call is

```
> noharm1d=noharm.sirt(pm=pmm,N=19601,dimensions=1,lower=0,
  reliability=T)
```

5. Performing more than a two-dimensional analysis with five items would not be meaningful. For example, with a three-dimensional solution one would have to allow items to cross-load; otherwise one factor would consist of only one item and would not be well defined. In addition, the number of independent parameters estimated with a three-dimensional solution would exceed the unique pieces of information (i.e., number of unique joint proportions).
6. When using NOHARM for dimensionality assessment, the results of the unidimensional solution also contain item parameter estimates. The model used for fitting the data is a two-parameter normal ogive model and is discussed in Appendix C. These item parameter estimates are not presented in Table 3.10. However, they are discussed in Appendix G, "Example: NOHARM Unidimensional Calibration." As a result, NOHARM can be used for both dimensionality assessment and item calibration for the one-, two-, and (modified) three-parameter models as well as their multidimensional versions.
7. The location estimates are biased under JMLE (Wright & Douglas, 1977). By averaging the relative bias over observed scores and items, Wright and Douglas (1977) arrived at a correction factor, $(L - 1)/L$, that can substantially reduce the bias

(assuming the distribution of item locations and of persons over observed scores is somewhat uniform). In BIGSTEPS and WINSTEPS, implementation of this correction factor is accomplished by setting STBIAS to “yes” (i.e., STBIAS = Y); the default for STBIAS is “no.” The effect of this correction is greatest for short instruments. Wright and Douglas (1977) suggest that for instruments of at least 30 items and item variability of about 1.0, there is little need for use of the correction factor.

8. As one removes either persons or items that have constant responses, it is possible to create additional response vectors that also have this characteristic for the remaining items and/or persons. For instance, assume that our data consist of the following binary response vectors from five persons on five items:

11111	← person 1
00000	← person 2
10111	← person 3
11110	← person 4
10000	← person 5

As we see, none of the *item response vectors* (i.e., the columns) have item scores of 0 or 5. However, the first two *person response vectors* (i.e., the rows) need to be deleted when using JMLE because person 1 has a perfect score of 5 and person 2 has a zero score. (Whenever one has zero variance for a binary response vector, it is not possible to use MLE for estimation.) After removing the first two cases, item 1’s item score is equal to 3 (i.e., the number of remaining cases) and is a zero-variance binary response vector. Upon removing item 1, the last case (person 5) has an observed score of 0 and needs to be removed. Removing the last case leaves two persons responding to the remaining four items. Inspecting the response vectors for these two people shows there are two items that have item scores equal to 2 and these two items need to be removed. Therefore, after culling all the zero-variance binary response vectors, the data consist of two persons (persons 3 and 4) responding to two items that have item scores of 1 and both persons have the same observed score ($X = 1$). Applying this logic to a perfect Guttman scale leads to the conclusion that a perfect Guttman scale cannot be calibrated using JMLE.

9. Guttman’s (1945) L_3 (i.e., λ_3) is identical to Cronbach’s (1951) α .
10. A variation of this approach, sometimes referred to as the “half-item rule,” has 0.5 assigned to the item with the smallest location value for a zero-score response vector and to the item with the largest location value for a perfect-score response vector. For example, assuming five items are in increasing order of their locations, then a zero-score response vector would become 0.5 0 0 0 0 and a perfect score vector would be 1 1 1 1 0.5. These modified vectors could then be used to obtain MLE $\hat{\theta}$ s for persons with zero and perfect scores.
11. We conjecture these 40 persons along with the 184 persons with $\underline{x} = 00001$ could account for items 1 and 5’s large OUTFIT ZSTD values. As verification, we remove these 224 persons and recalibrate the data. This results in the OUTFIT ZSTD value for item 1 becoming 0.8 and for item 5 falling to 0.5. The INFIT MNSQ value for item 1 decreases by 0.03 and increases by 0.03 for item 5; the INFIT MNSQ values

of the remaining items change by less than 0.02. For completeness, $\hat{\delta}_1 = -2.36$, $\hat{\delta}_2 = -0.26$, $\hat{\delta}_3 = 0.21$, $\hat{\delta}_4 = 1.04$, and $\hat{\delta}_5 = 1.36$, with a corresponding standard error for each item of 0.02. Therefore, there is a slight change in location estimates owing to the use of different people, but minimal impact on the accuracy of the estimates. In general, comparing calibration results with and without problem respondents can provide useful information for determining the best course of action.

12. Alternative splits that could be used for examining invariance and that might be of specific interest would be gender, racial/ethnic background (e.g., majority group individuals versus minority group individuals), educational level, high versus low on the θ continuum (e.g., high-proficiency versus low-proficiency individuals), and so on. Depending on the particular context, it may also make sense to examine estimate invariance over time. In each of these situations, we would perform separate calibrations for each group and then compare the results as we do in the examples in this chapter and in Chapter 5, or we would use a statistic, such as the likelihood ratio statistic (Andersen, 1973).

4

Marginal Maximum Likelihood Parameter Estimation

In this chapter, we present an alternative to JMLE that separates the estimation of item parameters from that of person parameters. This alternative, marginal maximum likelihood estimation (MMLE), estimates only item parameters.¹ As such, after obtaining item parameter estimates and achieving satisfactory model–data fit, one typically proceeds to estimate the person parameters using either MLE (see Appendix A) or a Bayesian approach. We begin this chapter with a conceptual presentation of MMLE, and then we discuss the Bayesian *expected a posteriori* method for estimating person locations. As we did in Chapter 3, we apply these methods to the data introduced in Chapter 2. Our fit analysis involves a graphical comparison of the empirical and predicted IRFs, with the items' empirical IRFs providing us with evidence supporting the tenability of the model's functional form assumption. We end by discussing the conversion of estimates from one metric to a different metric, and we introduce the concept of an instrument's characteristic function.

Marginal Maximum Likelihood Estimation

In JMLE one is simultaneously determining the item and person parameter estimates that maximize the joint likelihood of the observed data. This approach has a number of practical implications. First, there is the statistical issue of trying to simultaneously estimate the item parameters and the person parameters. Andersen (1970) referred to this issue as estimating *structural parameters* (the item parameters) concomitant with *incidental parameters* (the person parameters). The gist of this issue is that one way of obtaining better (e.g., consistent, efficient) item parameter estimates is to increase the sample size. However, in JMLE increasing the sample size leads to an increase in the number of person parameters to be estimated. Therefore, there are more person parameters to estimate without additional (item) information with which to estimate the person

parameters. Moreover, Neyman and Scott (1948) argue that when structural parameters are estimated simultaneously with the incidental parameters, the maximum likelihood estimates (MLEs) of the structural parameters do not have to be consistent as sample size increases. de Gruijter (1990) has shown that these estimates are biased.

Second, there is an issue of efficiency. Because the item and person parameter estimations are coupled together, if one finds that one or more items do not exhibit model–data fit and removes the item(s), then the instrument has to be recalibrated. The recalibration is necessary to remove the adverse effect(s) of the bad item(s) from the person location estimates as well as from the metric. Unfortunately, in JMLE one has to estimate the person locations as part of the process of estimating the item locations. It would be far more efficient to uncouple the item and person parameter estimation phases. In that case, estimation of the person parameters would not occur until one achieves satisfactory model–data fit.

Third, separating the person parameter estimation from the item parameter estimation may improve the theoretical accuracy of the estimates for some instruments. That is, for short instruments, say of 15 items or fewer, JMLE produces biased person location estimates that then result in poorly estimated item locations (Lord, 1983b, 1986).

MMLE provides an estimation approach that addresses these issues. (Other alternatives for handling the structural/incidental parameter issue include a Bayesian approach presented by Swaminathan and Gifford [1982] as well as by Jannarone, Yu, and Laughlin [1990], and for the 1PL/Rasch model the use of conditional maximum likelihood estimation.) In the following discussion, we begin with an analogy to two sampling situations in which the ANOVA framework may be applied. We then proceed to discuss MMLE.

To begin, consider two independent variables and the selection of their corresponding levels. In one case, the independent variables' levels may be the only levels of interest. For example, for the first independent variable the levels are males and females, and for the second the levels are treatment and control groups. These levels and their combination exhaust the set of levels for which we wish to make inferences. This situation may be analyzed using a fixed effects ANOVA model (see Hays, 1988; Scheffé, 1959).

Alternatively, for one factor the levels are randomly sampled from a population of levels, whereas the second independent variable's levels are the only ones of interest. For instance, consider a repeated measures design in which we randomly sample the participants, but the independent variable, say diet, consists of two diet regimens. In this case, we are interested in only making inferences for the specific treatment levels of the diet factor, but we want to be able to make inferences for the entire population of people from which the random sample is taken. The diet factor is a fixed effects factor, and the person factor is a random effects factor. This situation may be analyzed using a mixed effects ANOVA model.

In the IRT context, the fixed effects model is analogous to JMLE (Bock & Aitkin, 1981; Bock & Lieberman, 1970). In JMLE, the items are considered fixed because one is interested only in the particular items on the instrument and not in generalizing to other potential items that may have been included on the instrument. Persons are also considered fixed because it is only for the people in the calibration sample that one is interested in estimating person parameters. In contrast, Bock and Lieberman (1970)

proposed the MMLE solution by assuming that people are randomly sampled from a population (e.g., a normal distribution). In MMLE the items are still considered fixed, but the persons are considered a “random effect.” As such, the MMLE is analogous to a mixed effects ANOVA model.

In contrast to the ANOVA context where a random factor’s function is to allow the making of inferences from the sample to the population, in MMLE the random factor is a mechanism for introducing population information to estimation of the item parameters without having to directly estimate the person parameters. By invoking the idea that the calibration sample represents a random sample from a population, the estimation of the item parameters is free of the person location estimation. However, in this case, the estimation of the item parameters is potentially dependent on the population distribution.

The introduction of a population from which respondents are randomly sampled changes the estimation equations shown in Chapter 3. To arrive at the MMLE equations, we begin with the probability of the response vector \underline{x}

$$p(\underline{x} | \theta, \underline{\vartheta}) = \prod_{j=1}^L p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}, \quad (4.1)$$

where $p(\underline{x} | \theta, \underline{\vartheta})$ is conditional on the person location, θ , and on the item parameters ($\underline{\vartheta}$ is a matrix of item parameters [e.g., α and δ_j s]), and L is the instrument’s length.

Mathematically, application of the idea of randomly sampling respondents from a population to Equation 4.1 requires that one integrate over the population distribution

$$p(\underline{x}) = \int_{-\infty}^{\infty} p(\underline{x} | \theta, \underline{\vartheta}) g(\theta | \underline{v}) d\theta, \quad (4.2)$$

where $p(\underline{x} | \theta, \underline{\vartheta})$ is given by Equation 4.1, $g(\theta | \underline{v})$ represents the continuous population distribution of individuals, and \underline{v} is a vector containing location and population scale parameters that are typically 0 and 1, respectively (Thissen, 1982).² Note that $p(\underline{x})$ is not conditional on θ (i.e., the individual has been eliminated). Rather, Equation 4.2 specifies the unconditional or marginalized probability of a randomly selected person from the population, with a continuous latent distribution $g(\theta | \underline{v})$ providing the response vector \underline{x} (Baker, 1992; Bock & Aitkin, 1981). Therefore, although an individual’s θ is unknown, the probability of their possible θ s can be determined on the basis of their responses, the item parameters, and $g(\theta | \underline{v})$. Equation 4.2 is referred to as the marginal probability of the response vector \underline{x} . It is from this marginal distribution that the item parameters are estimated. In effect, in MMLE one multiplies the likelihood of the observed data by the population distribution to eliminate the person location parameters. Subsequently, one obtains MLEs of the item parameters by maximizing the resulting marginal likelihood function (cf. Lord, 1986; Thissen, 1982).

Because persons are eliminated from the estimation process, increasing the sample size does not increase the number of person parameters in the (marginal) likelihood function. The population distribution may be assumed to have a specific form *a priori*, such as a normal distribution, or it may be estimated empirically (Mislevy & Bock, 1997). It should be noted that IRT and JMLE do not make an assumption about the distribution of persons. However, we are making a distributional assumption with MMLE

if we assume *a priori* a particular distribution. Moreover, because this distributional assumption does not change from one individual to the next, all persons are considered to belong to the same distribution (cf. Appendix F, “Mixture Models”).

The gist of integration in Equation 4.2 is to determine the area under a function. This area corresponds to the probability of a person providing the response vector \underline{x} when the person is randomly selected from a population with a continuous latent distribution $g(\theta|\underline{y})$. Stated another way, to obtain this probability one needs to determine the area under the function used in Equation 4.2. Various approaches for performing numerical integration with computers have been developed. Some of these approaches use the Hermite–Gauss quadrature or are based on the Newton–Cotes integration formulas (e.g., the trapezoidal rule or Simpson’s rule). Because the technique typically associated with MMLE is the Hermite–Gauss quadrature method, this is the method we discuss.

The area defined by Equation 4.2 can be approximated by using a discrete distribution consisting of a series of rectangles (i.e., a histogram). The weighted sum of these rectangles provides an approximation to the area under the function (i.e., the probability). Obviously, for a given range on the continuum, say –4 to 4, the larger the number of rectangles, the narrower each rectangle becomes and the better the approximation. Each rectangle has a midpoint known as a *quadrature node* or *point* (X_r) and an associated *quadrature weight* ($A(X_r)$) that reflects the height of the function $g(\theta|\underline{y})$ around X_r . Figure 4.1 conceptually represents this idea for the unit normal distribution where a quadrature point is represented by a longish tick mark on the abscissa within each bar and the bar’s area reflects the quadrature weight; there are 11 quadrature points represented in the graph.³

Applying the quadrature points and weights to Equation 4.2 simplifies the calculation of the probability to a simple weighted sum of the conditional probabilities

$$p(\underline{x})^* \approx \sum_r^R p(\underline{x} | X_r, \underline{\vartheta}) A(X_r), \quad (4.3)$$

where the weighting is accomplished by using the quadrature weights, the sum is across the R quadrature points, and the probabilities are calculated using the quadrature points (X_r s) in lieu of θ . The quantity $p(\underline{x})^*$ is the unconditional probability of a randomly selected person from the population providing the response vector \underline{x} . Because Equation 4.3 uses quadrature nodes and weights, it is the “quadrature form” of Equation 4.2. As such, implementation of the Gaussian quadrature approximation results in the replacement of θ and the integral seen in Equation 4.2 by the quadrature points and a summation, respectively.

To apply the foregoing ideas to the entire $N \times L$ data matrix, we begin with the log likelihood of the marginalized likelihood function

$$\ln L = \sum_i^N \ln p(\underline{x}), \quad (4.4)$$

where $p(\underline{x})$ is given by Equation 4.2. By substitution and simplification, we obtain the

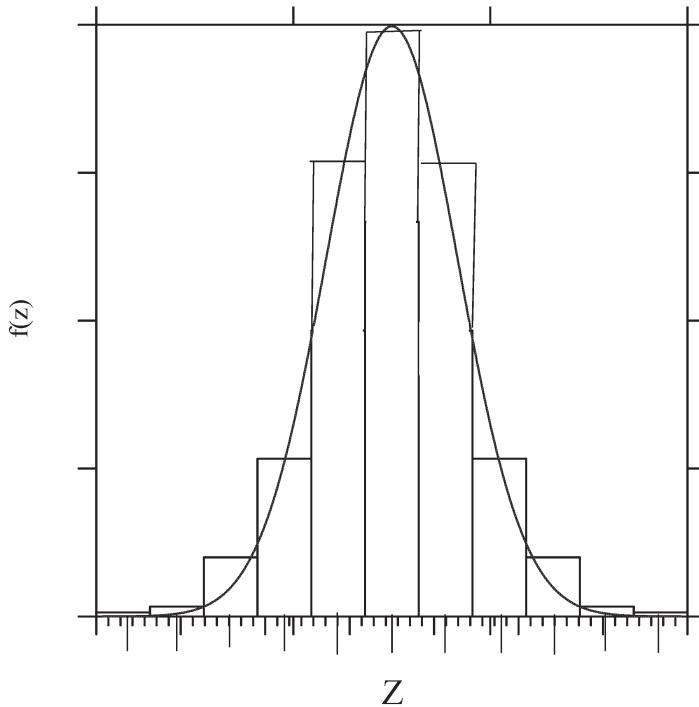


FIGURE 4.1. Conceptual representation of the quadrature concept with the unit normal distribution.

marginal likelihood equation to be solved for estimating an item's location based on individual response patterns

$$\frac{\partial}{\partial \delta_j} \ln L = - \sum_i^N \int \{ [x_{ij} - p_j(\theta_i)] [p_j(\theta_i | \underline{x}_i, \underline{\vartheta}, \underline{\nu})] \} d\theta = 0, \quad (4.5)$$

where $p_j(\theta_i | \underline{x}_i, \underline{\vartheta}, \underline{\nu})$ is given by

$$p_j(\theta_i | \underline{x}_i, \underline{\vartheta}, \underline{\nu}) = \frac{p(\underline{x}_i | \theta_i, \underline{\vartheta}) g(\theta_i | \underline{\nu})}{p(\underline{x}_i)}; \quad (4.6)$$

and for convenience we suppress the integration range of $-\infty$ to ∞ . The term $p(\underline{x}_i | \theta_i, \underline{\vartheta})$ is the likelihood function given by Equation 4.1, and $p(\underline{x}_i)$ is given by Equation 4.2. Conceptually, Equation 4.6 has the effect of spreading out the information provided by the response vector across the θ metric in direct proportion to the density of the population distribution (Baker, 1992; Harwell, Baker, & Zwarts, 1988).

Given the foregoing, Equations 4.5 and 4.6 may be rewritten in quadrature form. The quadrature form of Equation 4.5 is

$$\frac{\partial}{\partial \delta_j} \ln L = - \sum_r^R \sum_i^N \{ [x_{ij} - p_j(X_r)] [p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{\nu})] \} = 0 \quad (4.7)$$

and in quadrature form Equation 4.6 becomes

$$p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) = \frac{L(X_r)A(X_r)}{\sum_s^R L(X_s)A(X_s)}, \quad (4.8)$$

where

$$L(X_r) = \prod_{j=1}^L p_j(X_r)^{x_{ij}} (1 - p_j(X_r))^{(1-x_{ij})}.$$

The term $L(X_r)$ is the approximation of the likelihood function using the quadrature approach.

Equation 4.7 may be further simplified. First, expanding the product term in Equation 4.7 produces

$$\frac{\partial}{\partial \delta_j} \ln L = - \sum_r^R \sum_i^N \left\{ \left[x_{ij} * p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) \right] - \left[p_j(X_r) * p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) \right] \right\} = 0. \quad (4.9)$$

Second, distributing the summation across people results in

$$\frac{\partial}{\partial \delta_j} \ln L = - \sum_r^R \left\{ \left[\sum_i^N x_{ij} * p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) \right] - \left[p_j(X_r) \sum_i^N p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) \right] \right\} = 0. \quad (4.10)$$

Turning to the first term in Equation 4.10, let

$$\bar{c}_{rj} = \sum_i^N x_{ij} * p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}).$$

By substitution of Equation 4.8 into \bar{c}_{rj} we obtain

$$\bar{c}_{rj} = \sum_i^N x_{ij} * p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) = \sum_i^N \frac{x_{ij} L(X_r) A(X_r)}{\sum_s^R L(X_s) A(X_s)}. \quad (4.11)$$

Because \bar{c}_{rj} contains the binary response to item j , x_{ij} , the sum “counts” only the responses of 1 to item j for “persons” at X_r . As such, it reflects the expected number of responses of 1 to item j at each quadrature node X_r and is an *expected item score*.

Now looking at the second term in Equation 4.10, let

$$\bar{n}_{rj} = \sum_i^N p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}).$$

Again by substitution of Equation 4.8 we obtain

$$\bar{n}_{rj} = \sum_i^N p_j(X_r | \underline{x}_i, \underline{\vartheta}, \underline{v}) = \sum_i^N \frac{L(X_r) A(X_r)}{\sum_s^R L(X_s) A(X_s)}. \quad (4.12)$$

The term \bar{n}_{rj} is the *expected* number of persons at each quadrature point X_r . Equations

4.11 and 4.12 distribute the observed response vector for a person over the R quadrature nodes in proportion to the likelihood of the person being at the node (Mislevy & Bock, 1985).

Therefore, by substitution of Equations 4.11 and 4.12 into Equation 4.10, the marginal likelihood equation to be solved to estimate item j's location is

$$-\sum_r^R \left[\bar{c}_{rj} - \bar{n}_{rj} p_j(X_r) \right] = 0, \quad (4.13)$$

where $p_j(X_r)$ comes from the model using X_r in lieu of θ_i (Baker, 1992; Bock & Aitkin, 1981; Harwell, Baker, & Zwarts, 1988, 1989). Conceptually, the term $\bar{n}_{rj} p_j(X_r)$ says "of the persons expected to be at quadrature point X_r how many of them are expected to have a response of 1." Therefore, the difference in Equation 4.13 has the form of an expected "observed" item score minus an expected "predicted" item score. The value that minimizes the sum of these differences across the R quadrature points is the estimate of the item location, $\hat{\delta}_j$. (Quotation marks are used to distinguish the term *observed* from the way the term is used throughout this book [e.g., Appendix A]. The "observed" item score does not represent solely observed responses to item j but is based, in part, on the latent population distribution; an analogous interpretation can be provided for "predicted.") Examination of Equation 4.13 shows that it does not contain any reference to the θ s. Therefore, increasing the sample size does not lead to an increase in the number of parameters to be estimated; this has been true since the introduction of Equation 4.7.

As is the case with JMLE, estimating δ_j is an iterative process. The process begins with a provisional estimate of δ_j that is successively refined through a series of expectation and maximization cycles until Equation 4.13 is essentially 0.0 to a degree of accuracy determined by the convergence criterion. The Bock and Aitkin (1981) approach is an extension to a procedure developed for obtaining maximum likelihood estimates with probability models in the presence of incomplete data. This procedure is known as the *EM algorithm* (Dempster, Laird, & Rubin, 1977). In the current context, the person parameters are treated as the missing data.

The essence of the EM algorithm is to calculate \bar{n}_{rj} and \bar{c}_{rj} in the expectation step (i.e., the E-step). In the maximization step (i.e., the M-step), the values from the E-step are used to estimate the item parameters through an evaluation of the (marginal) likelihood function through a process similar to that discussed in Appendix B (i.e., find the maximum of this marginal likelihood function). Subsequent to the M-step, the refined M-step estimates are compared to the item parameter estimates used in the E-step. If the difference between these two sets of item parameter estimates is greater than the convergence criterion (and the maximum number of cycles has not been reached), then another E-M cycle is executed. Otherwise, convergence has been achieved and the estimation process is completed.

Because the EM algorithm does not provide sample standard errors, some programs (e.g., BILOG-MG) perform a series of Newton-Raphson (i.e., Newton-Gauss, Fisher-scoring) steps to improve the estimates and to obtain standard errors. Bock and Aitkin (1981) suggest stopping the EM steps short of convergence and using one or two New-

ton–Raphson iterations to improve the nearly converged EM solution and provide the standard errors of the item parameter estimates.

The \bar{c}_{rj} s and \bar{n}_{rj} s are typically referred to as “artificial data” or “pseudo-data” because they reflect the expected frequency and expected sample size, respectively, at each quadrature node (cf. Baker, 1992; Mislevy, 1986a; Thissen, 1982). The R \bar{n}_{rj} s represent the discrete posterior θ distribution. This posterior θ distribution is used to address the indeterminacy of scale by using the \bar{n}_{rj} s to adjust the quadrature weights from the previous E-step. Subsequently, the adjusted weights, $A(X_r)$, are normalized by dividing each expected number of persons at each quadrature point by the observed sample size (i.e., $A(X_r) = \bar{n}_{rj} / N$). These are then standardized (i.e., $\sum_r^R A(X_r)X_r = 0$ and $\sum_r^R A(X_r)X_r^2 = 1$) to set the location and scale of the latent variable (Baker, 1990; Mislevy & Bock, 1985). The process is repeated for each E-step. This strategy results in a “modified” person centering approach because it uses the posterior θ distribution rather than the distribution of individual $\hat{\theta}$ s. (The individual $\hat{\theta}$ s are not available at this point because of separation of the item parameter estimation from the person parameter estimation.) Baker (1992), Baker and Kim (2004), Bock and Aitkin (1981), Harwell and Baker (1991), Harwell et al. (1988, cf. Harwell et al., 1989), and Thissen (1982) provide greater detail about MMLE than is presented here. Moreover, Thissen’s (1982) presentation uses the frequency of response patterns rather than individual case data.

The use of a finite number of quadrature nodes is tantamount to assuming that these points are the only values that the θ s can take on (Mislevy & Stocking, 1989). Moreover, the statistical properties (e.g., consistent item parameter estimates) may not be realized if the assumed θ distribution is incorrect. A number of studies have investigated the effects of various factors on MMLE (e.g., Drasgow, 1989; Harwell & Janosky, 1991; Zwinderman & van der Wollenberg, 1990). For instance, Zwinderman and van der Wollenberg (1990) found that violation of the distributional assumption led to a loss of efficiency and biased estimates. Such bias and precision loss are directly related to the degree of violation. Seong (1990a) compared the item and proficiency parameter estimates from BILOG’s approach for determining the quadrature points and weights with those obtained by using the Stroud and Secrest values. Results showed that when a large number of quadrature points were used (e.g., 30, 40), the two methods estimated item and person location parameters equally well, but when a small number of quadrature points was specified (e.g., 10), the item and person parameter estimates were less accurate than when the Stroud and Secrest values were used. As is the case with MLE, it is not possible to estimate items that have zero variability in their binary response vectors (i.e., all responses to an item are 1 or are all 0).

Estimating an Individual’s Location: Expected A Posteriori

Application of MMLE yields only item parameter estimates. Once we are satisfied with our degree of model–data fit, we can proceed to obtain our person location estimates. There are several person estimation approaches that we can use. One estima-

tion approach is the MLE strategy introduced in Chapter 2 and covered in detail in Appendix A. However, as shown in Chapter 2, MLE cannot produce finite estimates of a person's location when he or she obtains either a zero score or a perfect score. However, in Chapter 2 we noted that additional information could reduce our uncertainty about a person's location. This additional information can come from previous experience or from assumptions. For example, if we believe or are willing to assume the construct of interest is normally distributed in the population, then the person could be considered to be sampled from a normal population. This in turn would provide information, in addition to the person's response vector, about where we would expect the person to be located. That is, assuming that one had a normal population, the probability of observing persons located between -1 and 1 would be more likely than, say, above 3 . A mechanism for incorporating this ancillary information into the estimation of person locations comes from Bayes' theorem.

The essence of this Bayesian strategy is that one has person location information in terms of a probability distribution prior to obtaining any observational data. This distribution is known as a *prior distribution*. After administering the instrument, one has observational data that is incorporated with the prior distribution information. The result of integrating the prior distribution with the observational data is a distribution referred to as the *posterior distribution* (i.e., because it comes after collecting the observations). This posterior distribution is used to obtain the estimate of a person's location. It should be noted the terms *prior* and *posterior* are relative (i.e., the posterior distribution can serve as the prior distribution in a second estimation cycle, and so on).

Figure 4.2 contains a conceptual representation of incorporating a prior distribution into the observational data (i.e., a likelihood function) to produce the posterior distribu-

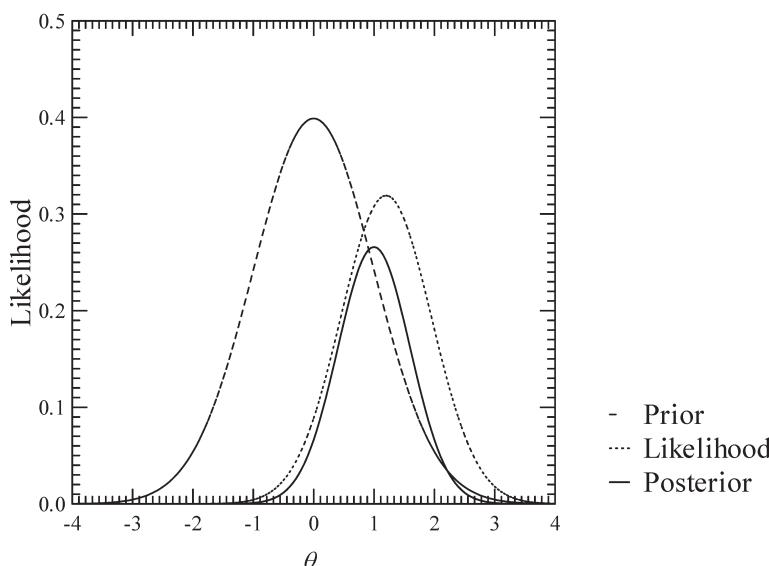


FIGURE 4.2. Conceptual presentation of the incorporation of a prior distribution into a likelihood function to produce the posterior distribution.

tion. In this case, we have the likelihood for the response pattern $\mathbf{x} = 110$. Assuming that a unit normal distribution is a reasonable prior to use, the result of incorporating this distribution information into the likelihood reduces the uncertainty about the person's location as reflected in the posterior distribution (i.e., the variability of the posterior is less than that of the prior). (The heights of the prior and posterior distribution are an artifact of how the plot is created.) Using this Bayesian strategy, we see that $\hat{\theta}$ would be approximately 0.90, the mean or mode of the posterior distribution.

If instead of using the likelihood function for $\mathbf{x} = 110$ we had used the likelihood function for, say a perfect score (e.g., Figure 2.8), then incorporating the prior distribution into this likelihood function would produce a posterior distribution that *had* a maximum. Consequently, the mean or mode could be used as $\hat{\theta}$. Other prior distributions are sometimes used in lieu of a normal distribution such as a beta distribution.

As mentioned, either the mean or the mode of the posterior distribution can be used to provide the person location estimate; in the case of a symmetric distribution both of these yield the same $\hat{\theta}$. In general, using the mean as the estimator minimizes the overall mean squared error of estimation provided that the appropriate prior is used. This mean squared error is less than that obtained by using MLE, although there is a concomitant increase in the estimation bias known as regression toward the prior's mean (Lord, 1986). One can get a crude sense of this regression by returning to Figure 4.2. Assume the person's true location is $\theta = 1.0$. The location of the maximum of L (dotted line) is the MLE $\hat{\theta}$, and it has an approximate value of 1.20. However, the Bayesian estimate of 0.90 represents the estimate being pulled toward the prior's mean of 0.0. (This is a conceptual representation of the regression toward the prior's mean phenomenon, and the value 0.9 is not the Bayesian estimate.) The estimation bias is variable and depends on the magnitude of the difference between the location of the prior and $\hat{\theta}$. The estimation bias is magnified the farther away $\hat{\theta}$ is from the prior's mean and minimized the closer $\hat{\theta}$ is to the prior's mean. The general pattern is one of underestimating large θ s and overestimating low θ s. Swaminathan and Gifford (1982) suggest that it is desirable to use priors that are neither too vague nor too specific. Selecting a somewhat diffuse prior mitigates to some extent the estimation bias.

Use of the mean or mode of the posterior distribution as an estimate corresponds to the two primary Bayesian strategies. In *maximum a posteriori* (MAP; also known as Bayes Modal Estimate), one uses the mode of the posterior distribution as the $\hat{\theta}$ (e.g., Swaminathan & Gifford, 1982), and in *expected a posteriori* (EAP, or Bayes Mean Estimate) one uses the mean of the posterior distribution as the $\hat{\theta}$ (Bock & Mislevy, 1982). The variability of the posterior distribution is used as an index of the error of estimation.

All three approaches for estimating a person's location (MLE, EAP, MAP) treat the item parameters' estimates as "known" and ignore their estimation error when estimating θ . Moreover, all three approaches can be modified to use *biweights* to reduce the sensitivity of a person location estimate to responses that are inconsistent with an IRT model (Mislevy & Bock, 1982).⁴ However, unlike MLE, both EAP and MAP can be used to obtain location estimates for all response patterns, including zero and perfect scores. Therefore, the advantage of a Bayesian approach over MLE is that finite $\hat{\theta}$ s are avail-

able for all individuals. Persons with perfect and zero scores have finite $\hat{\theta}$ s that are not artifacts dependent on an MLE correctional (kludge) system (e.g., as in BIGSTEPS). The disadvantage of a Bayesian approach is the (potential) bias that arises from the possible mismatch between an individual's θ and the prior's mean.

MAP differs from EAP in various ways. First, the MAP approach is an iterative method like MLE, whereas EAP is noniterative and based on numerical quadrature methods like those used in MMLE. Because of its noniterative (and efficient) nature, it is potentially faster than either MLE or MAP in estimating person locations. Second, MAP uses a continuous prior distribution, whereas EAP uses a discrete prior distribution. Third, whereas MAP $\hat{\theta}$ s exist for all response patterns, they suffer from greater regression toward the prior's mean than do the EAP estimates (Bock & Mislevy, 1982; Mislevy & Bock, 1997). Fourth, the average squared error for EAP estimates over the population is less than that for MAP (as well as MLE) person location estimates (Bock & Mislevy, 1982). A fifth difference has to do with implementation. Specifically, the mathematics required for deriving the computational forms for person location estimation with any IRT model are simpler with EAP than with MAP. In the following we discuss EAP; how to obtain MAP $\hat{\theta}$ s is shown in Chapter 8.

The EAP estimate (Bock & Mislevy, 1982) of an individual's θ after administering L items is given by

$$\hat{\theta}_i = \frac{\sum_{r=1}^R X_r L(X_r) A(X_r)}{\sum_{r=1}^R L(X_r) A(X_r)} \quad (4.14)$$

and its posterior standard deviation is

$$PSD(\hat{\theta}_i) = \sqrt{\frac{\sum_{r=1}^R (X_r - \hat{\theta}_i)^2 L(X_r) A(X_r)}{\sum_{r=1}^R L(X_r) A(X_r)}}, \quad (4.15)$$

where X_r , $A(X_r)$, and R are defined above and $L(X_r)$ is the likelihood function given the response pattern $\underline{x} = x_1, \dots, x_L$ and a particular model,

$$L(X_r) = \prod_{j=1}^J p_j(X_r)^{x_{rj}} (1 - p_j(X_r))^{(1-x_{rj})}. \quad (4.16)$$

As an example, we obtain the EAP $\hat{\theta}$ for $X = 2$ and $\underline{x} = 11000$ using 10 quadrature points and weights. Table 4.1 shows the relevant calculations. The likelihood function, $L(X_r)$, is obtained using the X_r s, the Rasch model, and the item parameter estimates ($\hat{\alpha} = 1.0$, $\hat{\delta}_1 = -2.155$, $\hat{\delta}_2 = -0.245, \dots, \hat{\delta}_5 = 1.211$); these estimates come from our example calibration below). Dividing the sum of column 4 by that of column 5 produces an EAP $\hat{\theta}$ for $X = 2$ of -0.2271 . The discrepancy between this $\hat{\theta}$ and each of the quadrature points is then used to modify the weighted likelihood to obtain $PSD(\hat{\theta})$. The $PSD(\hat{\theta})$ corresponding to our $\hat{\theta}$ is $(s_e(\hat{\theta})) 0.7290$. Use of the $PSD(\hat{\theta})$ as the standard error is based

on the fact that after 20 items the likelihood function and the posterior distribution are nearly identical and the $PSD(\hat{\theta})$ is virtually interchangeable with the standard error (Bock & Mislevy, 1982); the $PSD(\hat{\theta})$ s are labeled as standard errors in BILOG's EAP output.

Bock and Mislevy (1982) show that the EAP method produces reasonably accurate person location estimates. Seong (1990b) found that increasing the number of quadrature points from 10 to 20 produced more accurate $\hat{\theta}$ s, regardless of sample size and appropriateness of the prior distribution (i.e., normal, positively skewed, and negatively skewed). Given that person locations are estimated independently of one another, it is not surprising that sample size did not have a significant effect on the accuracy of EAP $\hat{\theta}$ s.

The accuracy of the EAP sample standard errors has been studied (e.g., de Ayala, Schafer, & Sava-Bolesti, 1995). The investigators found that using 10 quadrature points tended to result in EAP $s_e(\hat{\theta})$ s that underestimated the observed standard error. These $s_e(\hat{\theta})$ s give the false impression that $\hat{\theta}$ was being estimated more accurately than, in fact, it was. Therefore, if these $s_e(\hat{\theta})$ s are used to create confidence intervals (CIs), then the CIs would be erroneously narrower and classification decisions based on such CIs would potentially be incorrect. For these applications, it is recommended that one increase the number of quadrature points used in EAP estimation. A conservative approach would be to use 80 quadrature points because, overall, this level provides the best agreement

TABLE 4.1. EAP $\hat{\theta}$ for $x' = 11000$

X_r	$A(X_r)$	$L(X_r)$	$\hat{\theta}$		$PSD(\hat{\theta})$
			$X_r L(X_r) A(X_r)$	$L(X_r) A(X_r)$	
-4.0000	0.00012	0.0030369	-1.4456E-06	3.6139E-07	5.1444E-06
-3.1110	0.00281	0.0140101	-1.2226E-04	3.9298E-05	3.2684E-04
-2.2220	0.03002	0.0502904	-3.3546E-03	1.5097E-03	6.0082E-03
-1.3330	0.14580	0.1216368	-2.3640E-02	1.7735E-02	2.1690E-02
-0.4444	0.32130	0.1697289	-2.4235E-02	5.4534E-02	2.5753E-03
0.44440	0.32130	0.1178053	1.6821E-02	3.7851E-02	1.7067E-02
1.33300	0.14580	0.0382276	7.4296E-03	5.5736E-03	1.3565E-02
2.22200	0.03002	0.0064165	4.2801E-04	1.9262E-04	1.1554E-03
3.11100	0.00281	0.0006914	6.0338E-06	1.9395E-06	2.1612E-05
4.00000	0.00012	0.0000586	2.7884E-08	6.9711E-09	1.2456E-07
$\sum_{r=1}^R X_r L(X_r) A(X_r) = -0.0266689$					
$\hat{\theta} = \frac{-0.026689}{0.1174369} = -0.2271$					
$PSD(\hat{\theta}) = \sqrt{\frac{0.0624150}{0.1174369}} = 0.7290$					

between the CIs and their expected values. However, unless one needs to be concerned with the accuracy of standard errors, there does not appear to be a compelling reason to use 20 or more quadrature points, given that using 10 quadrature points provides very good agreement between the EAP $\hat{\theta}$ s and their corresponding θ s (i.e., $r_{\theta\hat{\theta}}$) for symmetric distributions. If there is reason to suspect that the latent person distribution is skewed, then use of $2\sqrt{L}$ quadrature points may be necessary (see Mislevy & Bock, 1997).

Example: Application of the Rasch Model to the Mathematics Data, MMLE, BILOG-MG

For this example, assume that we have already engaged in the various categories of activities (e.g., checking data integrity, assessing dimensionality) performed in Chapter 3, “Example: Application of the Rasch Model to the Mathematics Data, JMLE BIGSTEPS.” Although several programs are available for Rasch model estimation, such as flexMIRT, the R packages *mirt* (Chalmers, 2012, 2015, 2017, 2019) and *TAM*, and SAS *proc irt* (SAS Institute, 2012), in this example we use BILOG-MG as our calibration program; the terms *BILOG* and *BILOG-MG* are used interchangeably. (We use *proc irt* for a Rasch calibration in Chapter 13.) Subsequently, we reanalyze our data using the R package *mirt*. BILOG-MG’s fit statistic is a chi-square statistic. (Glas, 2007; Glas & Dagohoy, 2007; Glas & Verhelst, 1995a, 1995b; McKinley & Mills, 1985; Reise, 1990; and van den Wollenberg, 1988, contain other fit statistics that could be used.) As before, we use a graphical examination of fit by comparing empirical and predicted IRFs and use an item’s empirical IRF to determine whether our model’s functional form assumption is tenable for the item. This graphical examination complements the use of fit statistics.

BILOG-MG implements MMLE and exists on the Windows/Intel and compatible chips platforms. This program uses a GUI to create an input text command file. The GUI facilitates creating the command file by using a series of menus and dialogs to specify the data’s characteristics, the model to be used for calibration, the changing of default values, the nature of $g(\theta_i | \vartheta)$, and so on. After completing the menus and corresponding dialogs, one selects *Build Syntax* from the *Run* menu to create the command file. The command file (MATH.BLM) for our Rasch analysis is presented in Table 4.2; the data reside in a separate file called MATH.DAT. With BILOG-MG, we can analyze either individual case data or pattern data like that presented in Table 2.1. For this example we analyze individual case data.

The GLOBAL command line specifies the use of the logistic version (i.e., LOGistic) of the one-parameter model (i.e., NPARm = 1). This line, in conjunction with the CALIB command line’s use of the keyword RASch, specifies a Rasch model calibration.⁵ In addition, BILOG-MG is instructed to save the item parameter estimates to a file called MATHRSCH.PAR for future use (this requires the SAVE subcommand on the GLOBAL line and the file specification PARM = ‘MATHRSCH.PAR’ on the SAVE line). We spec-

TABLE 4.2. BILOG Command File for Rasch Calibration

```

Example Rasch Calibration

>GLOBAL DFName = 'C:\Math.dat',
   NPArm = 1,
   LOGistic,
   SAVE;
>SAVE PARm = 'MATHRSCH.PAR';
>LENGTH NITems = (5);
>INPUT NTotAl = 5,
   NIDchar = 10;
>ITEMS ;
>TEST1 TNAmE = 'TEST0001',
   INUmber = (1(1)5);
(10A1, T1, 5(1X,A1))
>CALIB ACCel = 1.0000,
   CHIsquare = (5, 8),
   PLOT = 1.0,
   RASch;

```

ify the total number of unique items (i.e., `NTotAl = 5`) and the use of 10 characters as an identification field (i.e., `NIDchar = 10`) on the `INPUT` line. For our example, we use the individual's response pattern as his or her identification field. We accomplish this by using the same first and last columns for the Case ID and Response String data fields in the Examinee Data . . . dialog's Data File tab (accessed from the Data menu). This dialog creates the FORTRAN format (i.e., [10A1,T1,5(1X,A1)]) used for reading the data. (On the FORTRAN format statement, the first data field [i.e., prior to the first comma] is the identification field [i.e., "10A1"], and the second data field [i.e., "5(1X,A1)"] reflects the responses. FORTRAN formats are discussed in Appendix G, "FORTRAN Formats.")

Although with fewer than 20 items the program's item fit chi-square statistics are unreliable, BILOG uses their probabilities to determine which empirical versus predicted IRFs to plot. Therefore, to obtain the IRF plots, we need to have BILOG calculate the item chi-squares. For our example, this is a two-step process. First, because by default BILOG-MG does not calculate the item chi-squares with fewer than 20 items, we need to tell BILOG to calculate the chi-square statistics with 5 items. We do this by inserting the subcommand `CHIsquare = (5, 8)` on the `CALIB` line; the 8 specifies the number of intervals to be used in calculating the chi-square value and will affect the appearance of the empirical IRF. Second, to have BILOG plot the empirical versus predicted IRFs for all the items, we increase the probability cutoff for display to 1 (i.e., `PLOT = 1.0` on the `CALIB` line). In the following, we ignore the calculated chi-squares. However, assuming that it is appropriate to use the chi-square statistic (e.g., $L \geq 20$), then this statistic tests the null hypothesis that an item's data are consistent with the model. Stated another way, we would like to see nonsignificant chi-square values. As is generally true, failure to reject the null hypothesis does not imply that the model is correct, but only that there is insufficient evidence to believe it is incorrect.

BILOG presents its output in three phases. Phase 1 results are found in a file with the same name as the command file, but with the extension PH1 (e.g., MATH.PH1). This phase contains information concerning the job setup, the reading of the data, and classical item statistics. Phase 2 contains the IRT calibration results, and Phase 3 contains person location estimates and related information.

Table 4.3 presents the BILOG-MG output for Phase 1. The Phase 1 output contains echoes of the commands shown in Table 4.2. The FILE ASSIGNMENT AND DISPOSITION section should be checked to ensure the correct data file is being used: SUBJECT DATA INPUT FILE MATH.DAT. (The additional listed [scratch] files [e.g., MF.DAT, CF.DAT, etc.] are created in the command file's folder and are automatically erased upon successful completion of the run. If the program "crashes," then these files will be littering the folder.) The line labeled ITEM RESPONSE MODEL indicates that the 1 PARAMETER LOGISTIC model and the logistic metric (i.e., LOGIT METRIC) are being used for the calibration.

In the DATA INPUT SPECIFICATIONS section we find details about the data to be calibrated. The number of items calibrated is indicated on the line labeled NUMBER OF ITEMS IN INPUT STREAM. The line labeled NUMBER OF SUBJECT ID CHARACTERS shows the identification field width as 10. (This value is the first field in the FORTRAN format; the format is echoed following the FORMAT FOR DATA INPUT IS line.) The TYPE OF DATA line indicates that individual case data are being used (i.e., SINGLE-SUBJECT DATA, NO CASE WEIGHTS) rather than, for example, pattern data. For this example all of these are correct. The first two observations are echoed (only the first one is presented), and their inspection shows that the data are read correctly.

On the line OBSERVATIONS READ FROM FILE . . . BILOG indicates the number of cases it read. This value should always be checked to verify that it matches the number of cases that are in the data file. Although there are many reasons why there might be a mismatch, two common problems are a misspecified format statement and/or an incorrect data filename. In our example, the correct number of cases is read from the correct file.

The ITEM STATISTICS section contains traditional item difficulty and discrimination statistics. If we had labeled the items, then their names would be found in the NAME column. Although knowing the proportion of people attempting an item is more useful than simply the number tried (#TRIED), the #TRIED column can provide information about potential speededness or instrument/person mismatch. For instance, omitted item(s) that occur at the end of an instrument may indicate that individuals were given insufficient time to respond; in this case, the instrument is said to be *speeded*.⁶ The ratio of the number right (#RIGHT) to #TRIED is presented in the percent column (labeled PCT). This column indicates the percentage of people correctly responding to an item. Dividing these percentages by 100% yields the traditional measure of an item's difficulty or *p*-value, P_j . In this case, we see that 88.7% of the examinees correctly answered item 1, and, as a consequence, it may be considered an "easy" item for this sample. In contrast, item 5 is comparatively more difficult for this sample because only 38.7% correctly answered it. The logit column is a transformation from the proportion difficulty (P_j) metric to a more "IRT-like" metric (logit = $\ln(P_j/(1 - P_j))$). To reconcile that the

TABLE 4.3. BILOG Output: Phase 1

```

:
>GLOBAL DFName = 'C:\Math.dat',
  NPArm = 1,
  Logistic,
  SAVe;

FILE ASSIGNMENT AND DISPOSITION
=====
SUBJECT DATA INPUT FILE      C:\MATH.DAT
BILOG-MG MASTER DATA FILE    MF.DAT      WILL BE CREATED FROM DATA FILE
CALIBRATION DATA FILE        CF.DAT      WILL BE CREATED FROM DATA FILE
ITEM PARAMETERS FILE         IF.DAT      WILL BE CREATED THIS RUN
CASE SCALE-SCORE FILE        SF.DAT
CASE WEIGHTING                NONE EMPLOYED
ITEM RESPONSE MODEL          1 PARAMETER LOGISTIC
                               LOGIT METRIC (I.E., D = 1.0)

>SAVE PARm = 'MATHRSCH.PAR';

BILOG-MG SAVE FILES
[OUTPUT FILES]
ITEM PARAMETERS FILE        MATHRSCH.PAR
:

DATA INPUT SPECIFICATIONS
=====
NUMBER OF FORMAT LINES           1
NUMBER OF ITEMS IN INPUT STREAM 5
NUMBER OF RESPONSE ALTERNATIVES 1000
NUMBER OF SUBJECT ID CHARACTERS 10
NUMBER OF GROUPS                 1
NUMBER OF TEST FORMS             1
TYPE OF DATA                     SINGLE-SUBJECT DATA, NO CASE WEIGHTS
MAXIMUM SAMPLE SIZE FOR ITEM CALIBRATION 10000000
ALL SUBJECTS INCLUDED IN RUN

:
FORMAT FOR DATA INPUT IS:
(10A1, T1, 5(1X,A1))

OBSERVATION #      1  WEIGHT:     1.0000  ID :  0 0 0 1 1
SUBTEST #:       1  TEST0001
  GROUP #:      1

  TRIED      RIGHT
    5.000     2.000
ITEM      1      2      3      4      5
TRIED     1.0     1.0     1.0     1.0     1.0
RIGHT     0.0     0.0     0.0     1.0     1.0
:
19601 OBSERVATIONS READ FROM FILE:   C:\MATH.DAT
:
ITEM STATISTICS FOR SUBTEST TEST0001
                                         ITEM*TEST CORRELATION
ITEM  NAME      #TRIED    #RIGHT    PCT    LOGIT    PEARSON    BISERIAL
-----
1    ITEM0001  19601.0  17395.0  88.7  -2.07    0.246    0.407
2    ITEM0002  19601.0  12624.0  64.4  -0.59    0.439    0.564
3    ITEM0003  19601.0  11094.0  56.6  -0.27    0.416    0.524
4    ITEM0004  19601.0  8369.0   42.7   0.29    0.405    0.511
5    ITEM0005  19601.0  7592.0   38.7   0.46    0.312    0.397
-----
```

interpretation of the traditional difficulty metric is the reverse of the IRT metric (e.g., P_j values close to 1.0 and large negative IRT logit values represent “easy” items), the logits are transformed to their opposite sign by multiplying them by -1.

The last two columns are collectively labeled ITEM*TEST CORRELATION and contain two traditional measures of item discrimination. The second to the last column (labeled PEARSON) contains the corrected point-biserial correlations, whereas the last column (labeled BISERIAL) contains the corresponding biserial correlations.⁷ If we use principal axis for dimensionality analysis and there are indications of more than one factor underlying the data, then inspecting the biserials may provide a clue to the presence of curvilinearity factors. According to McDonald (1985), “extremely sharply discriminating items” (p. 199) may lead to factors of curvilinearity (also known as “difficulty factors”).

Table 4.4 shows the Phase 2 results. The beginning of this output contains information about the execution: the maximum number of EM cycles (MAXIMUM NUMBER OF EM CYCLES: 20), Newton–Gauss cycles (MAXIMUM NUMBER OF NEWTON CYCLES: 2), the convergence criterion (CONVERGENCE CRITERION: 0.01), the assumption of a Gaussian person prior (LATENT DISTRIBUTION: NORMAL PRIOR FOR EACH GROUP), and the quadrature points and corresponding weights (e.g., the first quadrature node is $X_1 = -0.4000E+01 = -4.0$ and its weight $A(X_1) = 0.7648E-04 = 0.0007648$). Following this information is the calibration’s iteration history. Given that the last change across iterations (i.e., LARGEST CHANGE = 0.005) is less than the convergence criterion of 0.01 and that the number of executed EM and Newton cycles is less than their corresponding maxima, we have a converged solution; the number of EM cycles is 8 with 1 Newton cycle. The -2 LOG LIKELIHOOD values show the expected progressively decreasing pattern of a well-behaved solution. The marginal maximum log likelihood function value (-2 LOG LIKELIHOOD) after the last cycle may be used for comparing relative model fit; this is done in Chapter 6.

The table after the iteration history contains the item parameter estimates. From Chapter 2 we know the model may be written in a slope–intercept (i.e., linearized) form. In this form, the intercept, γ_j , is a function of an item’s discrimination and location parameters: $\gamma_j = -\alpha\delta_j$ (Equation 2.3). The INTERCEPT column contains the estimated item intercepts, γ_j s. The columns labeled SLOPE and THRESHOLD contain the fixed common discrimination parameter estimate, $\hat{\alpha}$, and item location parameter estimates, $\hat{\delta}_j$ s, respectively.⁸ The column labeled LOADING refers to the relationship between responses on item j and the (unidimensional) latent trait and is obtained by $\alpha_j/\sqrt{1+\alpha_j^2}$; see Appendix C and Appendix G, “Using Principal Axis for Estimating Item Discrimination.” The ASYMPTOTE column contains the estimates of the IRFs’ lower asymptote. This lower asymptote (χ) is associated with the three-parameter model and is discussed in Chapter 6. For the 1PL model $\chi_j = 0.0$ for all items.

The CHISQ and DF columns contain the chi-square statistics used for fit assessment and their degrees of freedom, respectively. These chi-square statistics are based on combining individuals into, by default, nine intervals on the basis of their Bayes location estimates (Mislevy & Bock, 1985). As mentioned above, we are ignoring these values because for our short instrument these values are suspect; their calculation is requested

TABLE 4.4. BILOG Output: Phase 2

```

:
      *** PHASE 2 ***
Example Rasch Calibration
>CALIB ACCel = 1.0000,
  CHIsquare = (5, 8),
  PLot = 1,
  RASch;

CALIBRATION PARAMETERS
=====
MAXIMUM NUMBER OF EM CYCLES:          20
MAXIMUM NUMBER OF NEWTON CYCLES:       2
CONVERGENCE CRITERION:                0.0100
ACCELERATION CONSTANT:                1.0000

LATENT DISTRIBUTION:                 NORMAL PRIOR FOR EACH GROUP
:
METHOD OF SOLUTION:
EM CYCLES (MAXIMUM OF    20)
FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF    2)

QUADRATURE POINTS AND PRIOR WEIGHTS:
      1      2      3      4      5
POINT -0.4000E+01 -0.3429E+01 -0.2857E+01 -0.2286E+01 -0.1714E+01
WEIGHT 0.7648E-04 0.6387E-03 0.3848E-02 0.1673E-01 0.5245E-01

      6      7      8      9      10
POINT -0.1143E+01 -0.5714E+00 -0.8882E-15 0.5714E+00 0.1143E+01
WEIGHT 0.1186E+00 0.1936E+00 0.2280E+00 0.1936E+00 0.1186E+00

      11     12     13     14     15
POINT 0.1714E+01 0.2286E+01 0.2857E+01 0.3429E+01 0.4000E+01
WEIGHT 0.5245E-01 0.1673E-01 0.3848E-02 0.6387E-03 0.7648E-04

[E-M CYCLES]
-2 LOG LIKELIHOOD = 112092.890

CYCLE 1; LARGEST CHANGE= 0.19046
-2 LOG LIKELIHOOD = 111293.340

CYCLE 2; LARGEST CHANGE= 0.12708
-2 LOG LIKELIHOOD = 110963.903

CYCLE 3; LARGEST CHANGE= 0.08050
-2 LOG LIKELIHOOD = 110842.600

CYCLE 4; LARGEST CHANGE= 0.015630
-2 LOG LIKELIHOOD = 110780.672

CYCLE 5; LARGEST CHANGE= 0.005751
-2 LOG LIKELIHOOD = 110778.567

CYCLE 6; LARGEST CHANGE= 0.007970
-2 LOG LIKELIHOOD = 110777.950

CYCLE 7; LARGEST CHANGE= 0.001872
-2 LOG LIKELIHOOD = 110774.876

CYCLE 8; LARGEST CHANGE= 0.000745
[FULL NEWTON CYCLES]
-2 LOG LIKELIHOOD: 110774.2948

CYCLE 9; LARGEST CHANGE= 0.000596

INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES
-----
0. 691. 3099. 4269. 4116. 4041. 3385. 0.
-----
INTERVAL AVERAGE THETAS
-----
***** -2.025 -1.293 -0.632 0.019 0.713 1.520*****
```

(continued)

TABLE 4.4. (*continued*)

SUBTEST TEST0001; ITEM PARAMETERS AFTER CYCLE 9									
ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF		
ITEM0001	2.155 0.028*	1.000 0.016*	-2.155 0.020*	0.707 0.012*	0.000 0.000*	160.2 (0.0000)	4.0		
ITEM0002	0.245 0.021*	1.000 0.016*	-0.245 0.015*	0.707 0.012*	0.000 0.000*	1285.5 (0.0000)	4.0		
ITEM0003	-0.206 0.020*	1.000 0.016*	0.206 0.014*	0.707 0.012*	0.000 0.000*	625.6 (0.0000)	4.0		
ITEM0004	-0.984 0.020*	1.000 0.016*	0.984 0.014*	0.707 0.012*	0.000 0.000*	367.6 (0.0000)	4.0		
ITEM0005	-1.211 0.020*	1.000 0.016*	1.211 0.014*	0.707 0.012*	0.000 0.000*	264.3 (0.0000)	4.0		

* STANDARD ERROR

LARGEST CHANGE = 0.005960

2702.9 20.0

(0.0000)

NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE
FOR TESTS WITH LESS THAN 20 ITEMS

PARAMETER MEAN STN DEV

THRESHOLD 0.000 1.340

QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

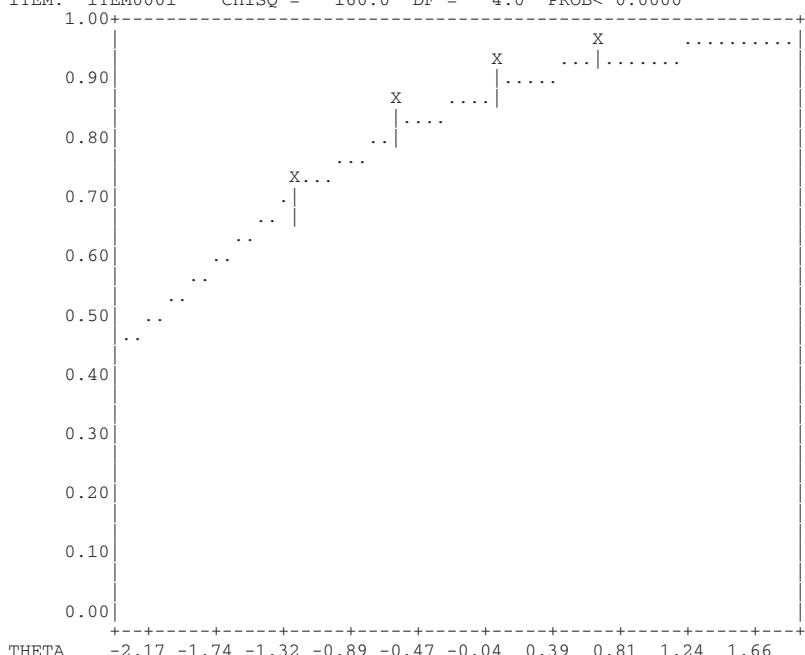
	1	2	3	4	5
POINT	-0.4057E+01	-0.3477E+01	-0.2898E+01	-0.2318E+01	-0.1739E+01
POSTERIOR	0.5977E-04	0.5148E-03	0.3272E-02	0.1537E-01	0.5209E-01

	6	7	8	9	10
POINT	-0.1159E+01	-0.5797E+00	-0.2245E-03	0.5793E+00	0.1159E+01
POSTERIOR	0.1228E+00	0.1979E+00	0.2246E+00	0.1890E+00	0.1184E+00

	11	12	13	14	15
POINT	0.1738E+01	0.2318E+01	0.2897E+01	0.3477E+01	0.4056E+01
POSTERIOR	0.5372E-01	0.1742E-01	0.4042E-02	0.6735E-03	0.7979E-04

MEAN 0.00000
S.D. 1.00000

ITEM: ITEM0001 CHISQ = 160.0 DF = 4.0 PROB< 0.0000



only as a means of obtaining the empirical versus predicted IRFs and for illustrative purposes.⁹

Given our Rasch model calibration, the common discrimination parameter estimate is set to 1.0 (i.e., $\hat{\alpha} = 1.0$). As can be seen from the THRESHOLD column, the item location estimates are $\hat{\delta}_1 = -2.155$, $\hat{\delta}_2 = -0.245$, $\hat{\delta}_3 = 0.206$, $\hat{\delta}_4 = 0.984$, and $\hat{\delta}_5 = 1.211$. The sample standard errors for these estimates are $s_e(\hat{\alpha}) = 0.016$, $s_e(\hat{\delta}_1) = 0.02$, $s_e(\hat{\delta}_2) = 0.015$, and 0.014 for $s_e(\hat{\delta}_3)$, $s_e(\hat{\delta}_4)$, and $s_e(\hat{\delta}_5)$. The mean and (inferential) standard deviation of the $\hat{\delta}$ s are 0.0 and 1.340, respectively. (If we use the IPL model for our calibration, then the $\hat{\alpha}$ column would potentially contain a value not equal to 1.0. Later in this chapter, in the “Metric Transformation and the Total Characteristic Function” section, we provide the IPL model estimates.)

Following the item parameter estimates table come the final adjusted quadrature weights and standardized quadrature points that determine the metric of the item parameter estimates. Comparing these adjusted weights and rescaled points with the initial $A(X_r)$ s and X_r , we see only slight changes in the weights.

Subsequent to the adjusted quadrature weights and points are the empirical versus predicted IRF plots. Only item 1's figure is presented in Table 4.4. (Prior to BILOG-MG, these IRFs plots were available only as the character-based graphics shown here. With BILOG-MG both character-based and bitmap graphics are available. A bitmap example of this type of plot is shown in Chapter 6, Figure 6.4.) The dotted line represents the model predicted IRF based on the item's parameter estimates. The Xs represent the proportion correct of a group of persons with approximately similar locations, whereas the vertical lines (error bands) represent a span of two standard errors around the expected group proportion tolerance interval (Mislevy & Bock, 1985). Collectively, these Xs are, in effect, the empirical IRF. If the Xs fall within the error bands, then there is agreement between the empirical IRF and the predicted IRF indicating item fit. For this particular item there is reasonable agreement, although for the other four items there is less agreement between the empirical and predicted IRFs. However, all empirical IRFs showed an ogival pattern consistent with the model's functional form assumption.

The degree of agreement between the empirical and predicted IRFs informs our judgment of fit but is not the sole determinant of our judgment. For instance, sometimes we find that the Xs reflect an ogival pattern that shows close agreement with the predicted IRF for a substantial range of the continuum, but disagreement at, for example, the lower end of the continuum (say, below -2). Depending on the application, this lack of fit at and below -2 may not be reason for concern. In short, different situations may be more amenable to or accepting of a certain degree of less than perfect fit. Another consideration is the number of intervals used in calculating the chi-square value and in creating the empirical IRFs. That is, as mentioned above, the number of intervals used can affect the appearance of these IRFs and, thereby, our interpretation. For example, with a small number of intervals we might observe strong agreement between the empirical and the predicted IRFs, but with a larger number of intervals the degree of agreement is not as strong, all other things being equal. Moreover, in making our judgment of fit, we recognize that the choice of two standard errors for defining the error bands is a reasonable, but arbitrary, decision. As a result, what defines agreement between the predicted

and empirical IRFs is not absolute. Again, all of this information is used to inform our judgment of fit along with the context (e.g., the number of items on the instrument, the number of items exhibiting “weak agreement,” the number of respondents, the purpose of the application).

Figure 4.3 contains a Double-Y graph for the total information function ($I(\theta)$) and the corresponding sample SEEs for the instrument. The shape of $I(\theta)$ is identical to that presented in Figure 2.11 (Chapter 2), although its maximum is located at different points along the θ continuum. This is because the two metrics are not aligned with one another (i.e., this is a result of the indeterminacy of metric issue). However, we can linearly transform the Figure 4.3 metric to align with that in Figure 2.11. As such, the total information function is invariant within a linear transformation; this is also true for the item information functions. From Equation 2.15 (Chapter 2) we know that there is an inverse relationship between $I(\theta)$ and SEE. We see this relationship in Figure 4.3. Specifically, as the total information (the solid line) for estimating θ decreases, the $s_e(\hat{\theta})$'s (the dashed line) for the $\hat{\theta}$'s increases, and vice versa. On this metric this instrument provides comparatively more information for estimating person locations between -0.5 and 1.25 than it does outside this range. Moreover, this figure may be used to determine the total information available (or standard error) for a specific θ . For example, we see that for individuals located at -3 , the total information (solid line) is about 0.34 (i.e., from the left ordinate scale) and that this corresponds to a standard error (dashed line) of approximately 1.72 (i.e., from the right ordinate scale).

Figure 4.4 presents the first item's IRF (the item's location estimate, $\hat{\delta}_1$, is identified by b on the abscissa; left panel) as well as its information function (right panel). As discussed in Chapter 2, for the 1PL model, the item's maximum information occurs at

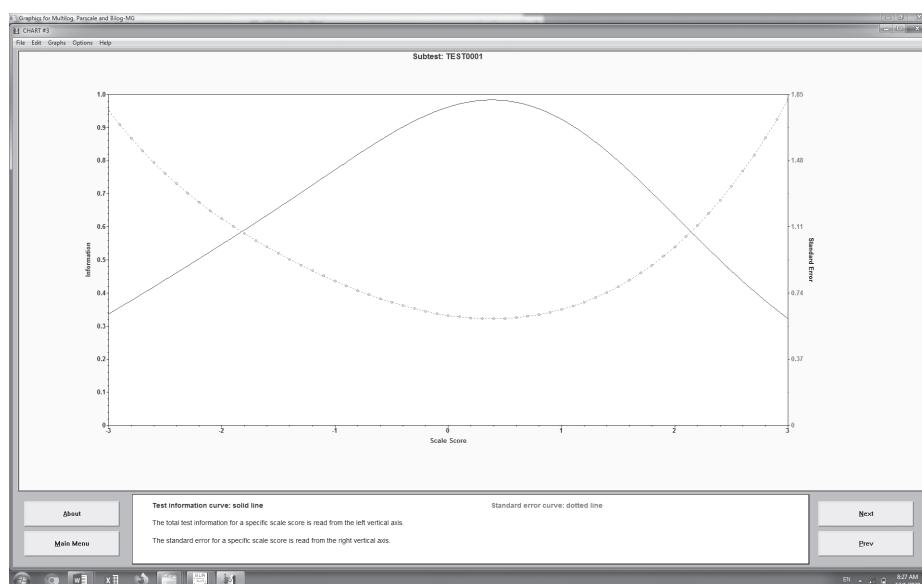


FIGURE 4.3. Total (test) information function (solid line) and standard error of estimate (dashed line).

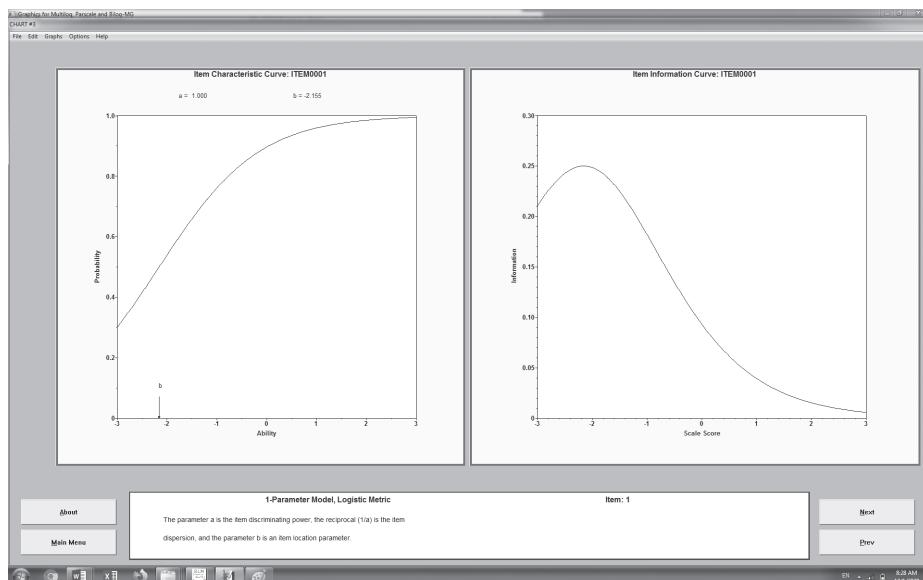


FIGURE 4.4. Item response and information functions for item 1.

an item's δ_j . Therefore, for the first item, its maximum information occurs at $\hat{\delta}_1 = -2.155$. If we present the $I(\theta)$ plot for the second item, then its maximum information would be the same as that for item 1, albeit at a different location (i.e., at $\hat{\delta}_2 = -0.245$). As would be expected and as seen from the right panel, the maximum item information for this item as well as for the other items on the instrument is 0.25.

As done in Chapter 3, we would also want to examine model–data fit by verifying the invariance of the estimates by splitting the sample into, say, two randomly equivalent groups and calibrating each group independently. (Other nonrandomly equivalent groups [e.g., males vs. females] may also be examined.) BILOG has the capability of sampling the data file by specifying SAMPLE on the INPUT line. The use of this feature for invariance assessment is demonstrated in Chapter 5.

In general, the constellation of item fit statistics, empirical versus predicted IRFs, and the tenability of assumptions need to be considered in determining model–data fit. From the foregoing we conclude that, practically speaking, there is an acceptable degree of fit for the instrument using the Rasch model. However, this does not mean that we cannot obtain better model–data fit with an alternative model and hence revisit these data using a different model in Chapter 5.

As previously mentioned, in BILOG the indeterminacy of the metric is addressed by imposing constraints on the estimated densities at the quadrature nodes to standardize $g(\theta_i | \underline{\delta})$ (Mislevy & Stocking, 1989). Because BILOG and BIGSTEPS address metric indeterminacy differently, we should not directly compare the $\hat{\delta}$ s from the calibrations without transforming one metric to the other. (Making comparisons across metrics is discussed later in this chapter.) However, theoretically and assuming model–data fit, the two δ metrics should be linearly related. Therefore, because the correlation is a standard-

ized linear relationship index (i.e., it is not tied to a particular metric) it is possible to use it to assess the relative linear ordered agreement across the metrics. Given our conclusion that there is model–data fit, we would expect the BILOG and BIGSTEP δ metrics to be highly linearly related. Comparing the BILOG estimates with those of BIGSTEPS (Table 3.4) shows that although they are not exactly equal, they are very close to one another. In fact, the Pearson correlation between the BILOG and BIGSTEPS' sets of $\hat{\delta}$ s is 0.9999. However, the variability of the BIGSTEPS $\hat{\delta}$ s is slightly larger ($SD = 1.3806$) than with the BILOG estimates. Another difference between the calibrations is BILOG's use of all 19,601 examinees and BIGSTEPS' use of 15,525; this is due to the programs' different estimation approaches (i.e., JMLE vs. MMLE).

Obtaining Person Location Estimates with BILOG-MG

Although we could obtain item and person parameter estimates in a single execution of BILOG, we separated the two stages for efficiency. That is, if the item fit analyses led to the removal of one or more item(s), then we would have had to re-execute BILOG on only the reduced item set. Separating the item parameter estimation stage from person parameter estimation allowed us to avoid estimating the locations for 19,601 persons until after we had achieved a satisfactory level of model–data fit. We now turn to estimating the person locations.

In estimating the person θ s, we do not re-estimate the item parameter estimates. Rather, we use our previously estimated items. Recall that the command file for estimating the item locations (Table 4.2) contained the SAVE subcommand on the GLOBAL command line and included the SAVe command (i.e., >SAVE PARM = 'MATHRSCH.PAR';). These keywords instruct BILOG to save the item parameter estimates to a file. It is this file's contents that we now use for obtaining the EAP person location estimates. Specifically, to estimate the person locations, we change the GLOBAL command line to read the item parameter estimate file by inserting the subcommand IFNAME = 'MATHRSCH.PAR'. Therefore, our GLOBAL command reads both the data and the item parameter estimate files (i.e., >GLOBAL DFName = 'C:\Math.dat',IFName = 'C:\MATHRSCH.PAR',NPA = 1,LOG;). In addition, we suppress item parameter estimation by replacing the CALIB line with >CALIB SELECT = 0; and we add the SCORE command with METHOD = 2 to specify (default) EAP estimation (i.e., >SCORE MET = 2;); MET = 2 is specified for pedagogical reasons.

Although the person location estimates are found in the third phase listing (e.g., MATH.PH3), we would normally want to inspect the Phase 1 listing to ensure that the item parameter estimate file is correctly identified and read. For example, we would check the FILE ASSIGNMENT AND DISPOSITION section for the line ITEM PARAMETERS FILE 'MATHRSCH.PAR,' as well as a line that reads PREVIOUSLY-PREPARED ITEM FILE READ FROM FILE: 'MATHRSCH.PAR'.

Table 4.5 contains part of our person location estimation results. On the line METHOD OF SCORING SUBJECTS: EXPECTATION A POSTERIORI, BILOG indicates that EAP estimation is performed. The section (GROUP SUBJECT IDENTIFICATION) contains the location estimates with two lines per person. The TRIED, RIGHT,

and PERCENT columns contain the number of items attempted, the number of responses coded 1, and the percent of 1s, respectively. The ABILITY column contains the $\hat{\theta}$ s, and the S.E. column the corresponding standard errors. As is the case in Chapter 3, we see that all individuals with the same observed score (i.e., the RIGHT column) obtain the same $\hat{\theta}$ regardless of the response pattern. Therefore, there are only six unique $\hat{\theta}$ s for all 19,601 persons. For example, compare the last three persons in the table with an $X = 4$. Because the individual's response pattern is used as his or her identification field, we see that for the first person with an $X = 4$ the response pattern is 01111. Furthermore, because the items are ordered by the δ_j s, we know that this individual incorrectly answered the easiest item and correctly answered the progressively more difficult items. The second of our three examinees had a response pattern of 10111, and the last person incorrectly responded to the hardest item but correctly answered all the easier items (i.e., $\mathbf{x} = 11110$). All three obtained a location estimate of 0.8238 with an $s_e(\hat{\theta})$ of 0.7292. It may also be noted that individuals with zero ($X = 0$) and perfect ($X = 5$) scores have $\hat{\theta}$ s of -1.3438 and 1.3685, respectively. Table 4.5 shows that the $\hat{\theta}$ for $X = 2$ agrees with our hand calculations above, as does the corresponding $s_e(\hat{\theta})$; see Table 4.1.

The last column in Table 4.5 contains the response vector's marginal probability (MARGINAL PROB). Above we defined the marginal probability as the probability of observing a response vector for a person randomly sampled from the population with a distribution of $g(\theta_i | \Psi)$. In the current context, the marginal probability is expressed in the denominators of Equations 4.14 and 4.15 (i.e., the MARGINAL PROB is given by $\sum L(X_r)A(X_r)$). Therefore, the individual's $\hat{\theta}$ is not used in calculating the probability of a particular response vector, \mathbf{x} . Rather, only the item parameter estimates and $g(\theta_i | \Psi)$, as approximated by using the quadrature points and weights X_r s and $A(X_r)$ s, are used for calculating the marginal probability of a response vector. The general pattern is that for a given observed score, the marginal probability is highest for the pattern that is consistent with intuition (i.e., an ideal response pattern) and decreases as the pattern becomes increasingly counterintuitive. For example, intuitively we would expect a person with an observed score of 3 to correctly answer the three easiest items (11100) rather than the three hardest items (00111). In Table 4.5 we see that a response vector of 11100 has a marginal probability of 0.099142, whereas the response vector of 00111 has a marginal probability of 0.001001. Therefore, given this instrument and the assumed distribution, the observed 11100 pattern is more probable than the 00111 response pattern. The other possible patterns with an $X = 3$ would have corresponding marginal probabilities between 0.099 and 0.001, with the values reflecting the degree of conformity to the 11100 pattern (e.g., 11010 would have a marginal probability greater than that for 11001).

Because the two calibration programs, BIGSTEPS and BILOG-MG, resolve the indeterminacy issue differently from one another, the MLE $\hat{\theta}$ s and the EAP $\hat{\theta}$ s are on different metrics and cannot be directly compared without a metric transformation. However, we can examine the relationship between the $\hat{\theta}$ s for two estimation approaches. The Pearson correlation between the MLE and the EAP $\hat{\theta}$ s corresponding to Xs of 1, 2, 3, 4 is 0.9978 and indicates the two metrics are highly linearly related.

TABLE 4.5. BILOG Output: Phase 3

:	PARAMETERS FOR SCORING, RESCALING, AND TEST AND ITEM INFORMATION							
METHOD OF SCORING SUBJECTS:	EXPECTATION A POSTERIORI (EAP; BAYES ESTIMATES)							
TYPE OF PRIOR:	NORMAL							
SCORES WRITTEN TO FILE	MATH.PH3							
TYPE OF RESCALING:	NONE REQUESTED							
ITEM AND TEST INFORMATION:	NONE REQUESTED							
DOMAIN SCORE ESTIMATION:	NONE REQUESTED							
:	EAP SUBJECT ESTIMATION, SUBTEST:MATH							
QUADRATURE POINTS AND PRIOR WEIGHTS, MEAN AND S.D.:								
POINT	1	2	3	4	5			
WEIGHT	-0.4000E+01	-0.3111E+01	-0.2222E+01	-0.1333E+01	-0.4444E+00			
POINT	0.1190E-03	0.2805E-02	0.3002E-01	0.1458E+00	0.3213E+00			
WEIGHT	0.4444E+00	0.1333E+01	0.2222E+01	0.3111E+01	0.4000E+01			
WEIGHT	0.3213E+00	0.1458E+00	0.3002E-01	0.2805E-02	0.1190E-03			
MEAN	0.0000							
S.D.	1.0000							
GROUP	SUBJECT IDENTIFICATION							
WEIGHT	TEST	TRIED	RIGHT	PERCENT	ABILITY	S.E.	MARGINAL PROB	
-----	-----	-----	-----	-----	-----	-----	-----	-----
1	1 1 0 0 0							
1.00	TEST0001	5	2	40.00	-0.2272	0.7291	0.117440	
1	1 0 0 0 0							
1.00	TEST0001	5	1	20.00	-0.7698	0.7457	0.150947	
1	1 1 1 0 0							
1.00	TEST0001	5	3	60.00	0.2984	0.7228	0.099142	
1	1 0 0 1 1							
1.00	TEST0001	5	3	60.00	0.2984	0.7228	0.010609	
1	0 0 1 1 1							
1.00	TEST0001	5	3	60.00	0.2984	0.7228	0.001001	
1	0 0 0 0 0							
1.00	TEST0001	5	0	0.00	-1.3438	0.7706	0.050159	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.064910	
1	1 1 1 1 1							
1.00	TEST0001	5	5	100.00	1.3685	0.7491	0.057713	
1	0 1 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.002240	
1	1 0 1 1 1							
1.00	TEST0001	5	4	80.00	0.8238	0.7292	0.015124	
1	1 1 1 1 0							

Metric Transformation and the Total Characteristic Function

This chapter and the preceding chapters have described situations where we compared estimates across different metrics; for example, we directly compared the BILOG estimates with those of BIGSTEPS or directly compared estimates from different subsamples. Short of just making statements about the linear agreement of estimates across the two metrics, we could not directly compare the estimates from the two different metrics because of (potential) differences in the origins and units used. As a consequence, to make these comparisons we need to align the two metrics with one another to produce a common metric.

Although this idea is more fully examined in Chapter 11, we present the essence of it here. Typically, one chooses one metric to serve as the *target* metric onto which the other metric is transformed. This is analogous to choosing the Fahrenheit temperature scale as the target metric and transforming Celsius temperatures to Fahrenheit. (As is also true with the temperature scales in this analogy, the IRT metric is relative, not absolute.)

As previously mentioned, our continuum is determined up to a linear transformation. In general, the linear transformation from one metric to another metric for both person and item locations is given by

$$\xi^* = \zeta(\xi) + \kappa, \quad (4.17)$$

where ξ is the δ_j (or θ_i) on the metric to be transformed (the *initial* metric) and ξ^* is the δ_j^* (or θ_i^*) on the target metric. Collectively, ζ and κ are called the *metric transformation coefficients* or *equating coefficients*. (In the Celsius/Fahrenheit analogy $\zeta = 1.8$ and $\kappa = 32$.)

To transform our other item parameter, item discrimination, we would use

$$\alpha^* = \frac{\alpha}{\zeta}. \quad (4.18)$$

In terms of a slope–intercept parameterization, our intercept parameter is transformed by

$$\gamma^* = \gamma - \frac{\alpha(\kappa)}{\zeta}. \quad (4.19)$$

In some situations, the values of ζ and κ are given by the target metric's characteristics. For instance, if we are interested in converting our θ s (or their estimates) to the T-score scale to enhance their interpretability, then the target metric is the T-score scale and, by definition, $\zeta = 10$ and $\kappa = 50$. In other situations, we might wish to align the metric from one calibration sample with that of another sample or align the Rasch calibration metric with a 1PL model calibration of the same data. (We refer to this alignment of metrics as *linking* and discuss it further in Chapter 11.) In these cases, we may need to estimate the values of ζ and κ .

Multiple strategies can be used to obtain the metric transformation coefficients.

One simple approach is based on using the means and standard deviations of the item locations. In this approach, the transformation coefficient ζ is obtained by taking the ratio of the target to initial metric item location standard deviations

$$\zeta = \frac{s_{\delta}^*}{s_{\delta}}, \quad (4.20)$$

where s_{δ}^* is the standard deviation of the item locations on the target metric and s_{δ} is the standard deviation of the item locations on the initial metric. Once ζ is determined, the other transformation coefficient κ is obtained by

$$\kappa = \bar{\delta}_j^* - \zeta \bar{\delta}_j, \quad (4.21)$$

where $\bar{\delta}_j^*$ is the mean of the item locations on the target metric and $\bar{\delta}_j$ is the mean of the item locations on the initial metric. Equations 4.20 and 4.21 are the standard linear transformation equations one sees in an introductory statistics course. (An alternative approach for determining ζ and κ is discussed in Chapter 11.)

As an example, assume we wish to link the metric from our Rasch model calibration with that from a 1PL model calibration of the mathematics data. To obtain the 1PL model calibration estimates, we remove the keyword RASch from the CALIB command line (Table 4.2) and re-execute BILOG. Our 1PL model item parameter estimates are $\hat{\alpha} = 1.421$, $\hat{\delta}_1 = -1.925$, $\hat{\delta}_2 = -0.581$, $\hat{\delta}_3 = -0.264$, $\hat{\delta}_4 = 0.284$, and $\hat{\delta}_5 = 0.443$. The mean and standard deviation of the $\hat{\delta}$ s are -0.409 and 0.943 , respectively. Table 4.4 shows that our Rasch model estimates are $\hat{\alpha} = 1.0$, $\hat{\delta}_1 = -2.155$, $\hat{\delta}_2 = -0.245$, $\hat{\delta}_3 = 0.206$, $\hat{\delta}_4 = 0.984$, and $\hat{\delta}_5 = 1.211$ with a mean $\hat{\delta}$ of 0.0 . Comparing the two sets of estimates shows that the two metrics are not the same, although they are highly linearly related ($r = 0.9999$).

To transform the Rasch model calibration metric (i.e., the initial metric) to that of the 1PL model calibration metric (i.e., the target metric), we apply Equations 4.17 and 4.18. We obtain our metric transformation coefficients by using the metrics' means and standard deviations. Given that the 1PL and Rasch models' $\hat{\delta}$ s standard deviations are 0.943 and 1.340 , respectively, we have

$$\zeta = \frac{s_{\delta}^*}{s_{\delta}} = \frac{0.943}{1.340} = 0.704.$$

Because the respective initial and target metric means are 0.0 and -0.409 , we have that

$$\kappa = \bar{\delta}_j^* - \zeta \bar{\delta}_j = -0.409 - 0.704(0) = -0.409.$$

Therefore, the transformation equation for item (and person) location estimates is

$$\hat{\xi}^* = \zeta(\hat{\xi}) + \kappa = (0.704) \hat{\xi} + (-0.409)$$

and for the item discrimination parameter we use

$$\alpha^* = \frac{\alpha}{0.704}.$$

For example, to transform the Rasch $\hat{\alpha}$ of 1.0 to the target (1PL model) metric, we have

$$\alpha^* = \frac{1.0}{0.704} = 1.421$$

Transforming the location estimate for item 1 to the target metric yields

$$\hat{\delta}_1^* = \zeta(\hat{\delta}_1) + \kappa = 0.704(\hat{\delta}_1) + (-0.409) = 0.704(-2.155) - 0.409 = -1.9257.$$

For the other items, the transformed estimates are $\hat{\delta}_2^* = -0.5814$, $\hat{\delta}_3^* = -0.2640$, $\hat{\delta}_4^* = 0.2835$, and $\hat{\delta}_5^* = 0.4433$.

Of course, the correlation between the two metrics is still 0.9999. Now that we have aligned our metrics, we can directly compare our 1PL model and Rasch model estimates to one another. As would be expected, given the mathematical equivalency of the Rasch and 1PL models, the two sets of estimates are the same. (Any differences in the transformed estimates and the target values are due to item parameter estimation error as well as rounding error.) Comparing our results, we see that the effect of having forced α to be 1.0 is to stretch out the metric relative to when α is estimated to be 1.421. Because α 's value is absorbed into the metric, we can stretch or contract the metric by changing the value of α .

As mentioned above, another use of a metric transformation is to convert a metric to make it more meaningful or interpretable. Focusing on the person location estimates, we could convert our standard θ metric that is centered at 0.0 to a target metric that did not have negative values (e.g., a T-score scale, the College Entrance Examination Board [CEEB] scale). We would do this by using Equation 4.17, with ξ representing θ and the appropriate values for ζ and κ .

Another target metric that has intrinsic meaning for people is the total score metric. For instance, rather than informing a respondent that their $\hat{\theta}$ is 1.1746, which may or may not have any inherent meaning to the person, we can transform the individual's $\hat{\theta}$ to the more familiar total score metric.

We can perform this transformation through the *total characteristic function* (TCF)

$$T = \sum_{j=1}^L p_j, \quad (4.22)$$

where T is the expected trait score, L represents the instrument's length, and p_j is the probability of a response of 1 according to a dichotomous IRT model; Equation 4.22 is also called the *test characteristic function*. In a proficiency assessment situation, the total score metric indicates the number of expected correctly answered items.

From Equation 4.22 we see that θ and T represent the same concept, but on different metrics (Lord, 1980). That is, with θ we have an infinite metric $-\infty < \theta < \infty$, whereas with T and the 1PL model we have a bounded metric, $0 \leq T \leq L$. As a result, when $\theta = -\infty$, T equals 0, and when $\theta = \infty$, then T equals the number of items on the instrument (a perfect score). However, θ and T differ in that T 's metric is dependent on the items on the instrument and the θ 's metric is independent of the instrument's items.

In addition, because the relationship between p_j and θ is nonlinear, the relationship between θ and T is also nonlinear.

In some cases, it may be desirable to convert θ to a proportion metric. To obtain an expected proportion equivalence for θ (i.e., the proportion of responses of 1), we divide T by L

$$\mathbf{ET} = \frac{\sum p_j}{L}. \quad (4.23)$$

The term \mathbf{ET} is referred to as the *expected proportion trait score* or the *domain score* (e.g., see Hambleton & Swaminathan, 1985; Lord, 1980). In proficiency assessment, \mathbf{ET} is the expected proportion of correct responses.

As an example, assume that we wish to report performance on our mathematics test on the total score metric (i.e., 0 to 5), but we also want to have the advantages that IRT provides over CTT. As a result, rather than reporting observed scores, we calibrate our data with the 1PL model and estimate our examinees' locations. To convert these $\hat{\theta}$ s to the total score metric we use Equation 4.22. For instance, for individuals located at 1.1746 (i.e., $\hat{\theta} = 1.1746$), we calculate the probabilities of a response of 1 for each of our items. These probabilities are $p_1 = 0.98793$, $p_2 = 0.92377$, $p_3 = 0.88537$, $p_4 = 0.77998$, and $p_5 = 0.73877$. Therefore, all examinees with a location estimate of 1.1746 have a trait score of

$$T = 0.98793 + \dots + 0.73877 = 4.31581.$$

or, in terms of expected proportion correct, we have

$$\mathbf{ET} = \frac{4.31581}{5} = 0.8631.$$

That is, any person with a location estimate of 1.1746 would be expected to correctly answer 4.32 items, or 86% of the items, on our mathematics examination.

We can graphically represent the functional relationship between θ and T (i.e., Equation 4.22). This graphical relationship is the *total characteristic curve* (TCC) or *test characteristic curve*. For example, Figure 4.5 contains the TCC for our five-item mathematics instrument and shows the transformation of 1.1746 to its corresponding T. We can easily convert any $\hat{\theta}$ to its corresponding T by simply identifying the person's location on the abscissa, projecting from this point up to the TCC, and then proceeding to the ordinate. Moreover, the TCC graphical representation allows us to take any T and convert it to its corresponding θ .

From the foregoing we see that the total characteristic function not only specifies the relationship between the trait score T and the IRT person location (or its estimate), but also facilitates the conversion of θ (or $\hat{\theta}$) to the total score metric. This function has one additional use. In Chapter 11 we use it to place different calibration results onto a common metric. This approach, known as *total characteristic function equating*, provides an additional method for obtaining our metric transformation coefficient values.

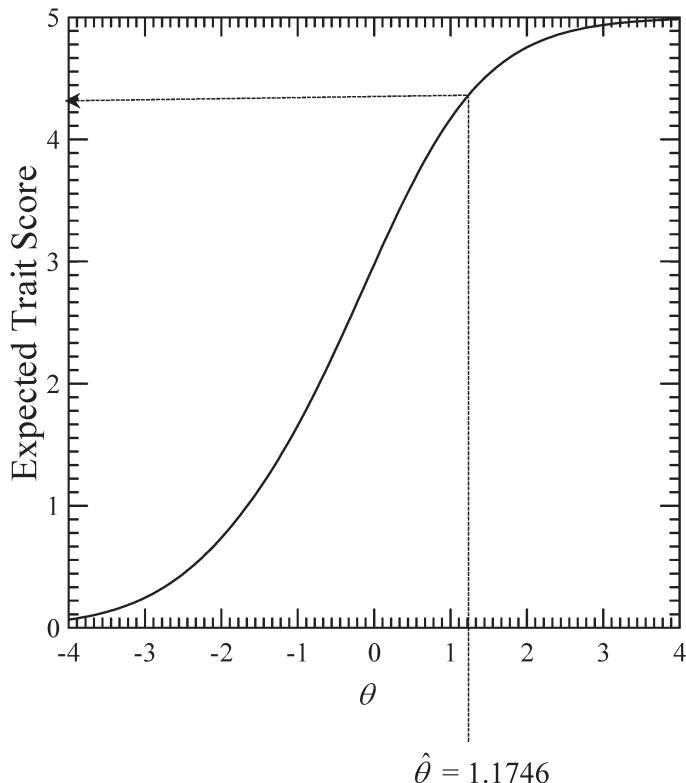


FIGURE 4.5. Total characteristic curve.

Example: Application of the Rasch Model to the Mathematics Data, MMLE, mirt

We demonstrate the calibration of our data using an R package, and we select `mirt`—a flexible package for fitting dichotomous and polytomous models as well as multidimensional models. It offers multiple fit statistics, simulation capability, differential item functioning analysis (Chapter 12), and imputation of missing data, to highlight just a few capabilities. Table 4.6 shows our R session. Instead of using the `read.table` function with the data filename specified as we did in Chapter 3, we use the `file.choose()` function (`read.table(file.choose())`) to open a file dialog window from which we select our data file `Math.dat` (i.e., interactive file selection). The data are brought into our workspace and stored in the object (i.e., a data frame) `mathdata`. To ensure that the correct number of cases was read and that the data were read correctly, we examine the first five and last five cases (`head(mathdata, 5)` and `tail(mathdata, 5)`, respectively); we actually compare these responses to the corresponding cases in `Math.dat`. Using the `summary` function, we obtain descriptive statistics for our items. For example, our `min = 0` and `max = 1` show that we do not have any nonbinary responses. Moreover, item 1's traditional difficulty (P_1) is 0.8875, for item 2 we have $P_2 = 0.644$, and so on.

TABLE 4.6. mirt Session for the Rasch Calibration of the Mathematics Data

```

> library(mirt)

> # read data & provide meaningful variable names
> mathdata = read.table(file.choose(), col.names=c(paste0("I", 1:5)))

> head(mathdata, 5) # show the first 5 cases of the data file
  I1 I2 I3 I4 I5
1  1  1  0  0  0
2  1  1  1  0  0
3  1  0  0  0  0
4  1  1  1  0  0
5  1  0  1  1  0

> tail(mathdata, 5) # show the last 5 cases of the data file
  I1 I2 I3 I4 I5
19597 1  1  1  1  0
19598 1  1  1  1  0
19599 1  1  1  1  1
19600 1  1  0  1  1
19601 1  1  1  1  0

> summary(mathdata)
      I1          I2          I3          I4          I5
Min. :0.0000  Min. :0.0000  Min. :0.0000  Min. :0.0000  Min. :0.0000
1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
Median :1.0000  Median :1.0000  Median :1.0000  Median :0.0000  Median :0.0000
Mean   :0.8875  Mean   :0.644   Mean   :0.566   Mean   :0.427   Mean   :0.3873
3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000

> rasch = mirt(mathdata, 1, 'Rasch', SE=T, SE.type='Fisher')
  Iteration: 38, Log-Lik: -55387.029, Max-Change: 0.00009

  Calculating information matrix.

> rasch
  Call:
  mirt(data = mathdata, model = 1, itemtype = "Rasch", SE = T,
    SE.type = "Fisher")

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 38 EM iterations.
  mirt version: 1.30
  M-step optimizer: nlmminb
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Information matrix estimated with method: Fisher
  Condition number of information matrix = 8.723652
  Second-order test: model is a possible local maximum # see Endnote 10

```

(continued)

TABLE 4.6. (continued)

```

Log-likelihood = -55387.03
Estimated parameters: 6
AIC = 110786.1; AICC = 110786.1
BIC = 110833.4; SABIC = 110814.3
G2 (25) = 772.4, p = 0
RMSEA = 0.039, CFI = NaN, TLI = NaN

> M2(rasch,CI=0.95) # Maydeu-Olivares & Joe statistic
      M2 df p    RMSEA   RMSEA_2.5 RMSEA_97.5   SRMSR      TLI      CFI
stats 428.241  9 0 0.04875089 0.04411739 0.05349489 0.04211215 0.9597483 0.9637735

> itemfit(rasch,fit_stats='infit',which.items=1:5)
  item outfit z.outfit infit z.infit
1  I1  0.675 -10.222 0.991 -0.477
2  I2  0.615 -37.398 0.728 -40.891
3  I3  0.658 -40.760 0.742 -41.453
4  I4  0.674 -37.789 0.736 -41.778
5  I5  0.763 -23.764 0.816 -27.323

> itemfit(rasch,group.bins=6,empirical.plot=2) # produces Figure 4.6

> # Item parameter estimates for all 19,601 cases
> coef(rasch,IRTpars=TRUE,printSE=T)
$I1
  a     b     g     u
par 1 -2.736 0 1
SE NA 0.019 NA NA

$I2
  a     b     g     u
par 1 -0.826 0 1
SE NA 0.017 NA NA

$I3
  a     b     g     u
par 1 -0.375 0 1
SE NA 0.018 NA NA

$I4
  a     b     g     u
par 1 0.403 0 1
SE NA 0.021 NA NA

$I5
  a     b     g     u
par 1 0.630 0 1
SE NA 0.022 NA NA

$GroupPars
  MEAN_1 COV_11
par 0 2.073
SE NA 0.011

```

(continued)

TABLE 4.6. (continued)

```

> # examination of invariance -----
> set.seed(50000)
> caseU=runif(19601)
> sortmathdata=mathdata
> sortmathdata$unif=caseU
> sortmathdata=sortmathdata[order(sortmathdata$unif), ] # 19601 x 5 variables
# sort data by random numbers

> mathdata1=sortmathdata[1:9800, ] # 9800 x 6 variables
> mathdata2=sortmathdata[9801:19601, ] # 9801 x 6 variables

> mathdata1=within(mathdata1,rm(unif)) # drop unif var; 9800 x 5
> mathdata2=within(mathdata2,rm(unif)) # drop unif var; 9801 x 5

> # subsample 1 calibration
> rasch1 = mirt(mathdata1,1,'Rasch',SE=T,SE.type='Fisher')
  Iteration: 38, Log-Lik: -27731.126, Max-Change: 0.00009

  Calculating information matrix...

> coef(rasch1,simplify=TRUE,IRTpars=TRUE)
  $items
    a   b   g   u
  I1 1 -2.746 0 1
  I2 1 -0.837 0 1
  I3 1 -0.380 0 1
  I4 1  0.381 0 1
  I5 1  0.645 0 1

  $means
  F1
  0

  $cov
    F1
  F1 2.092

> # subsample 2 calibration
> rasch2 = mirt(mathdata2,1,'Rasch',SE=T,SE.type='Fisher')
  Iteration: 38, Log-Lik: -27652.120, Max-Change: 0.00008

  Calculating information matrix...

> coef(rasch2,simplify=TRUE,IRTpars=TRUE)
  $items
    a   b   g   u
  I1 1 -2.727 0 1
  I2 1 -0.815 0 1
  I3 1 -0.370 0 1
  I4 1  0.424 0 1
  I5 1  0.615 0 1

  $means
  F1
  0

```

(continued)

TABLE 4.6. (continued)

```

$cov
  F1
F1 2.055

> cor(s1_itest$b,s2_itest$b)          # see Endnote 14
[1] 0.9998197

> # subsample 1
> mean(s1_itest$b)
[1] -0.5873126
> sd(s1_itest$b)
[1] 1.343829

> # subsample 2
> mean(s2_itest$b)
[1] -0.5747106
> sd(s2_itest$b)
[1] 1.336455

# CI overlap as check for invariance -----
# subsample 1
> print((s1_itest=coef(rasch1,IRTpars=TRUE)))
$I1
      a     b     g     u
par    1 -2.746  0   1
CI_2.5 NA -2.799 NA NA
CI_97.5 NA -2.694 NA NA

$I2
      a     b     g     u
par    1 -0.837  0   1
CI_2.5 NA -0.883 NA NA
CI_97.5 NA -0.791 NA NA
:

> # subsample 2      (metrics not aligned)
> print((s2_itest=coef(rasch2,IRTpars=TRUE)))
$I1
      a     b     g     u
par    1 -2.727  0   1
CI_2.5 NA -2.779 NA NA
CI_97.5 NA -2.675 NA NA

$I2
      a     b     g     u
par    1 -0.815  0   1
CI_2.5 NA -0.861 NA NA
CI_97.5 NA -0.769 NA NA
:

> # subsample 2 (metric aligned to subsample 1)
> s2_iteststar # see Endnote 15
      loc    CI_2.5    CI_97.5
1 -2.7394381 -2.7914591 -2.6874171
2 -0.8279822 -0.8740567 -0.7819078

```

(continued)

TABLE 4.6. (continued)

```

3 -0.3830117 -0.4315253 -0.3344982
4  0.4110409  0.3540264  0.4680553
5  0.6028284  0.5430610  0.6625959

> # subsample 1's estimates & CIs
> s1_mtrx
      loc    CI_2.5    CI_97.5
1 -2.7462109 -2.7985071 -2.6939147
2 -0.8369125 -0.8830959 -0.7907291
3 -0.3799882 -0.4286372 -0.3313393
4  0.3813749  0.3247025  0.4380473
5  0.6451739  0.5847205  0.7056273

> # end of invariance check results ---
> itemplot(rasch,1,"infotrace" , theta_lim=c(-4,4))          # produces Figure 4.7 left graph
> itemplot(rasch,2,"infotrace" , theta_lim=c(-4,4))          # produces Figure 4.7 left graph

> # obtaining person estimates via fscores & displaying first 6 cases
> head((peopleRasch=fscores(rasch,method="EAP", full.scores.SE=T)),6)
      F1     SE_F1
[1,] -0.69603477 0.8589810
[2,]  0.03519389 0.8576724
[3,] -1.46119268 0.8959412
[4,]  0.03519389 0.8576724
[5,]  0.03519389 0.8576724
[6,]  0.03519389 0.8576724

> tail(peopleRasch,4)                                         # display last 4 cases
      F1     SE_F1
[19598,] 0.8003611 0.8998282
[19599,] 1.6913664 0.9970738
[19600,] 0.8003611 0.8998282
[19601,] 0.8003611 0.8998282

> mean(peopleRasch[,1])                                       # average person location
[1] -1.713367e-05
> sd(peopleRasch[,1])                                         # SD of person location
[1] 1.122373

> # obtaining person fit info via personfit & display first 6 cases
> head((peopleRaschFit=personfit(rasch,method="EAP")), 6)
      outfit  z.outfit   infit  z.infit      Zh
1 0.4664704 -0.8999909 0.5435514 -1.1218155  0.9978743
2 0.4785387 -0.5671475 0.5586181 -1.4093791  1.1019782
3 0.2850797 -1.0983718 0.3268722 -1.3071571  1.0500414
4 0.4785387 -0.5671475 0.5586181 -1.4093791  1.1019782
5 1.0175206  0.2846527 1.1612633  0.5574013 -0.3314355
6 0.4785387 -0.5671475 0.5586181 -1.4093791  1.1019782

> tail(peopleRaschFit,4)  # display last 4 cases
      outfit  z.outfit   infit  z.infit      Zh
19598 0.4783240 -0.1267654 0.6467544 -0.7260871  0.7116548
19599 0.1681778 -0.4790043 0.2424252 -1.0105332  0.8121423
19600 0.9961875  0.4046188 1.1787052  0.5202511 -0.2930955
19601 0.4783240 -0.1267654 0.6467544 -0.7260871  0.7116548

```

^aWe use the set.seed function to allow the reader to generate the same set of random numbers as the example. Consequently, this function is optional.

We call the `mirt` function and store its output in the object `rasch`. In our call to `mirt`, we specify our data frame (object), `mathdata`, the dimensionality for the calibration (i.e., 1), and the model to estimate, ‘Rasch,’ and we indicate that we want the Fisher information matrix calculated (`SE = T`, `SE.type = ‘Fisher’`) for obtaining the standard errors. The data frame passed to `mirt` contains only the data to be calibrated. Therefore, any variables in the data frame that are not to be calibrated (e.g., a case identification variable) need to be removed using the `remove` function `rm()`; Chapter 5 shows an example of using `rm()`. Although we invoke `mirt` in a simple fashion, the use of additional arguments would allow one to take advantage of more of its capabilities (e.g., specification of the estimation algorithm, number of quadrature points, survey weights).

To display the contents of the `rasch` object, we simply type the output object’s name. Although the output object contains all the calibration results, by default typing the object’s name will display program execution parameters (e.g., optimizer used, number of quadrature points) and model-level fit information. The default maximum number of EM iterations is 500, with a convergence criterion of 0.0001. Our calibration required only 38 iterations to obtain convergence (i.e., `Converged within 1e-04 tolerance after 38 EM iterations`). Our $-2\ln L$ value is $-2(-55387.03) = 110,774.06$, with information criteria values of $AIC = 110,786.1$ and $BIC = 110,833.4$. (As above, we ignore the information criteria until Chapter 5.)

To obtain additional model-level fit information, we pass our output object to the `M2` function. The `M2` function calculates M_2 (Maydeu-Olivares & Joe, 2006) along with RMSEA, SRMR, TLI, and CFI. (These indices are discussed further in Appendix G, “CFI, GFI, M_2 , RMSEA, TLI, and SRMR.”) In our call, `M2(rasch, CI = 0.95)`, we request the RMSEA’s 95% confidence interval (CI). (The 95% CI indicates the accuracy of our RMSEA point estimate, with the ideal case having the CI, including 0.) Using a 5% significance level, we see that $M_2 = 428.241$ is significant, and we can reject the model–data fit null hypothesis. In contrast, our other fit indices indicate model–data fit.

The RMSEA’s 95% CI [RMSEA $_2.5 = 0.04412$ to RMSEA $_97.5 = 0.05350$] is comparatively narrow, and we can expect the true value to be within this range 95% of the time; $RMSEA = 0.04875$. Thus, our 95% CI indicates that we are in the “close fit” to “good” fit neighborhood. That is, according to guidelines, a RMSEA less than 0.05 indicates a “close fit” and a value from 0.05 to 0.08 a “good/fair” fit. (Because greater than 0.08 is indicative of “poor” fit, we want our CI’s upper bound to be less than 0.08.) Because our SRMR of 0.04211 is less than the “close to” cutoff of 0.08, it reflects good fit. Moreover, our TLI (a.k.a., NNFI) and CFI values ($TLI = 0.95975$, $CFI = 0.96377$) also reflect good fit ($TLI \geq 0.95$; $CFI \geq 0.95$). In short, we have contradictory information. However, our sample size provides M_2 with a great deal of power to reject the null hypothesis of model–data fit. As a result, we do not give M_2 as much weight as our other indices, and we interpret our RMSEA 95% CI, SRMR, TLI, and CFI as providing evidence of model–data fit.

We examine item-level fit by using the `itemfit` function to produce the `INFIT` and `OUTFIT` values for our items. Specifically, to obtain these fit statistics, we specify the output object, the fit statistics to calculate (`fit.stats = ‘infit’`), and we want to have them for all five items (`which.items = 1:5`). Recall that values around 1 are

considered good, with values substantially different from 1 indicating either dependency or noise. Using the guideline for a “run of the mill” instrument (Wright & Linacre, 1994) of 0.7 to 1.3 for delineating “acceptable,” we see that our infit values fall within this range. Although items 1–4’s outfit values are close to the lower bound, they fall below it. Fortunately, these low values indicate data overfit to the Rasch model and are “less of a threat to measurement than INFIT” (Linacre, 2002). As above, we ignore the statistical tests `z.infitz` and `outfit` due to our large N. (Because MMLE allows us to use all our examinees, these infit and outfit values will not match those in Table 3.4.)

`mirt`’s `itemfit` function can produce a plot to compare an item’s empirical IRF with its predicted IRF. Although with only five items such a plot would not be particularly useful for pedagogical reasons, we demonstrate obtaining one. For example, a call of the `itemfit` function (`itemfit(rasch, group.bins = 6, empirical.plot = 2)`) specifies our output object, to place examinees into six approximately equal-size groups (“bins”) and to produce the empirical and predicted IRFs for item 2 (Figure 4.6).¹¹ (We do not add CIs about our empirical points [e.g., using the `empirical.CI = .95` argument] because the width of a CI is, in part, a function of the N and our

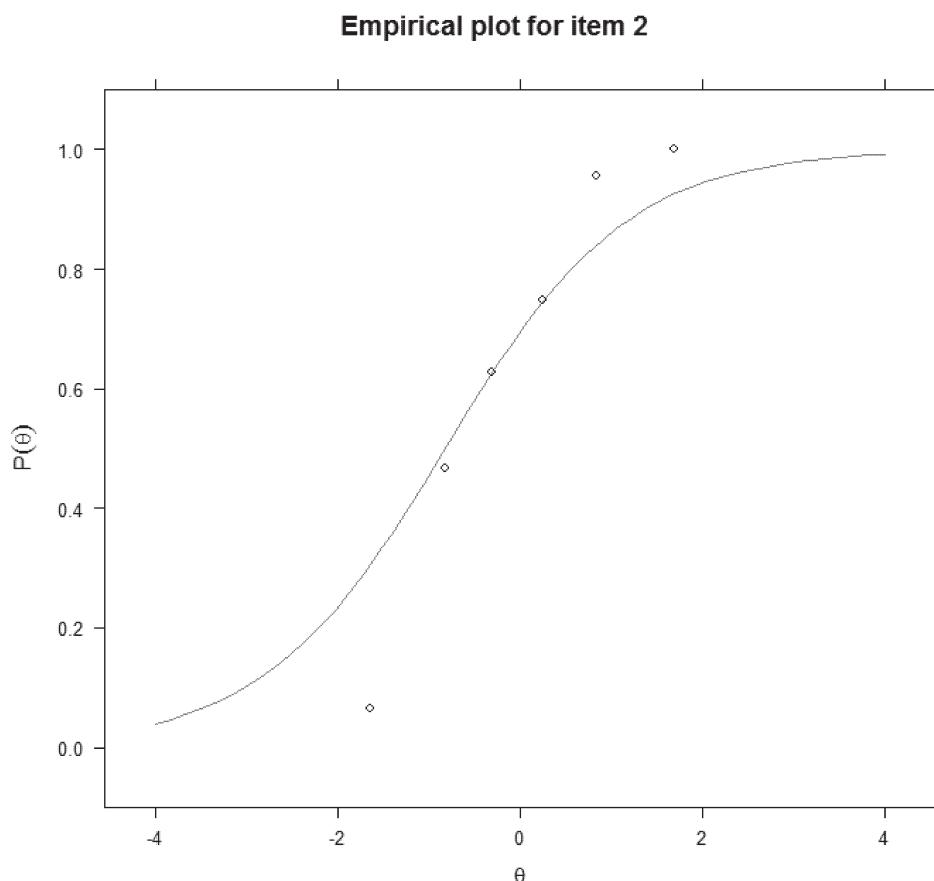


FIGURE 4.6. IRF for item 2 with observed proportions.

sample size is so large.)¹² We specify six fractiles (“group bins”) because we have six unique $\hat{\theta}$ s; the number of fractiles typically alters the appearance of the empirical IRF. For the lowest group (i.e., below -0.9), we see that we expect about five times as many correct responses as we observed. As we move up the continuum, there is greater agreement between what we expect given the Rasch model and what we observed; this item is located at -0.826 . At the upper end we observed slightly more correct responses than we would predict, but the discrepancy is not as large as with the lowest bin. Overall, we have good correspondence between the empirical and predicted IRFs between approximately -0.9 and 0.30 , with relatively good correspondence above 0.30 . Because we may not wish to weight the lowest and highest fractiles as much as those that are toward the center, we do not have a fit concern with this item.¹³ In addition, we have evidence supporting the functional form assumption. (If we were interested in modeling the observed data, then increasing the slope of the IRF will improve the empirical-predicted IRF agreement.)

We use the `coef()` function with our output object (i.e., `coef(rasch, IRTpars = T, printSE = T)`) to obtain our item parameter estimates and their standard errors. By default, our item location estimates are on an easiness scale ($\hat{\delta}_j^E$), not a difficulty scale, ($\hat{\delta}_j$) where $\hat{\delta}_j = -\hat{\delta}_j^E$. We can override this by setting `IRTpars = TRUE`. The column labeled ‘b’ contains our item locations ($\hat{\delta}_1 = -2.736$, $\hat{\delta}_2 = -0.826$, $\hat{\delta}_3 = -0.375$, $\hat{\delta}_4 = 0.403$, and $\hat{\delta}_5 = 0.630$); the column is labeled ‘d’ when the item locations are on an easiness scale. The sample standard errors for these estimates are $s_e(\hat{\delta}_1) = 0.019$, $s_e(\hat{\delta}_2) = 0.017$, $s_e(\hat{\delta}_3) = 0.018$, $s_e(\hat{\delta}_4) = 0.021$, and 0.022 for $s_e(\hat{\delta}_5)$. The column labeled ‘a’ contains the item discrimination value (i.e., $\alpha = 1$); the columns labeled ‘g’ and ‘u’ are the IRF’s lower (χ_j) and upper (Υ_j) asymptotes, respectively. (For the Rasch/IPL model $g = 0$ and $u = 1$.) Our person population has a M of 0 (`MEAN = 1`), and an estimated variance (`COV = 11`) is 2.073 .

Given the above BILOG results, it is natural to compare our `mirt` estimates to those of BILOG. Comparing these `mirt` estimates with those shown in Table 4.4, we see that the same issue discussed above in regards to BIGSTEPS’s estimates applies here. As was the case between BIGSTEPS’s and BILOG’s estimates, the correlation between these item location estimates and those of BILOG is $r = 1.000$. Moreover, our location estimate scale is not aligned with that of the BILOG estimates; BILOG mean $\hat{\delta} = 0.0002$ and `mirt` mean $\hat{\delta} = -0.5808$. By subtracting the difference between the mean $\hat{\delta}_j$ s from each of the `mirt` $\hat{\delta}$ s we align the `mirt` scale with that of BILOG. The result is that the BILOG and `mirt` $\hat{\delta}$ s are found to be essentially the same.

In an examination of item parameter invariance, we divide our sample into two random subsamples. To create our subsamples, we generate a uniform random number ($[0, 1]$) for each of our examinees (`caseU = runif(19601)`), duplicate our data frame (`sortmathdata = mathdata`), add our random number to `sortmathdata` (`sortmathdata$unif = caseU`), and sort `sortmathdata` into the ascending order of the random number (`sortmathdata [order(sortmathdata$unif),]`). Our first subsample (`mathdata1`) is the first 9800 examinees, with the second (`mathdata2`) containing the remaining examinees; we remove the `unif` variable from both subsamples’ data frames. Each subsample is in turn calibrated with their results stored in the output objects `rasch1` and `rasch2`.

The two sets of location estimates show a strong linear relationship ($r = 0.99982$) with

the corresponding scatterplot (not shown) indicating that the correlation captures the linearity pattern in the data; `plot(s1_itest$b, s2_itest$b, xlab = "RS 1", ylab = "RS 2")`). Accordingly, we have evidence of item parameter invariance.

An alternative to the point estimates approach used with the correlation is to use CIs. With CIs, evidence of item parameter invariance is overlapping CIs for corresponding items across our subsamples. Because our estimates' Ms are different (subsample 1: $\hat{\delta} = -0.58731$; subsample 2: $\hat{\delta} = -0.57471$), we need to align our metrics before comparing our item parameter estimates and/or their CIs. We transform subsample 2's metric to align with subsample 1's (target) metric. Because we are using the Rasch model, we only need be concerned with the item location estimates. After aligning our metrics, we compare our subsamples CIs.^{14,15} Because each of subsample 2's CIs overlap with the corresponding subsample 1 CIs, we have evidence of item parameter estimate invariance. As an example, subsample 2 item 1's transformed CI of $[-2.7916, -2.6874]$ (see `subsample 2 (metrics aligned)`) overlaps with subsample 1 item 1's CI $[-2.7985, -2.6939]$. Thus, we have evidence of estimate invariance for item 1. Of course, if our second subsample's $\hat{\delta}$ s fall within the corresponding subsample 1 item's CI (as they do in this example), we can forego examining corresponding CIs.

For convenience we examine our items' predicted IRF and item information in a single graph by using the `itemplot` function and the `infotrace` type. The `itemplot` function will also produce an item's predicted IRF (`trace`), information (`info`), and standard error (`SE`), to name just a few. Our call produces a Double-Y graph, with the left ordinate representing the probability scale and the right ordinate providing the scale for the item information function. As examples, we request the item information and IRFs for items 1 and 2 (e.g., for item 1: `itemplot(rasch, 1, "infotrace", theta_lim = c(-4,4))`). As would be expected with the Rasch model, the maximum item information is the same for these two items, albeit located at different points, $\hat{\delta}_1$ and $\hat{\delta}_2$ (Figure 4.7). (See Endnote 16 for how to perform a 1PL calibration and Endnote 17 for how to obtain the TCF.)

Given the model- and item-level fit, we proceed to estimating our person proficiencies by using the `fscores` function. Our call (`fscores(rasch, method = "EAP", full.scores.SE = T)`) requests the EAP $\hat{\theta}$ s and their standard errors, and stores the results in the `peopleRasch` object; because EAP is the default estimation approach, we could have simply written `fscores(rasch, full.scores.SE = T)`. We embed our call of `fscores` within a call to the `head` function to display the first six $\hat{\theta}$ s. We use the `tail` function with our output object to display the last four $\hat{\theta}$ s. For our example, $\hat{\theta}_1 = -0.69604$ ($s_e(\hat{\theta}_1) = 0.85898$), $\hat{\theta}_2 = 0.03520$ ($s_e(\hat{\theta}_2) = 0.85767$), and so on. Although our M and SD are different from BILOG's, our estimates correlate with BILOG's 0.99950 (`mirt: M = -0.00002` and `SD = 1.12237`, BILOG: $M = 0.25026$ and $SD = 0.76683$). As above, we can transform our `mirt` $\hat{\theta}$ s to be on the same metric as BILOG's. Performing this transformation results in the mean discrepancy between the two sets of $\hat{\theta}$ s to be 0.00000. Our unique $\hat{\theta}$ s for X s of 0, 1, 2, 3, 4, and 5 are -2.32289 , -1.46119 , -0.69603 , 0.03519 , 0.80036 , and 1.69137 , respectively. We can obtain person fit statistics (`personfit(rasch, method = "EAP")`). Our discussion in Chapter 3 on interpreting these values applies here.

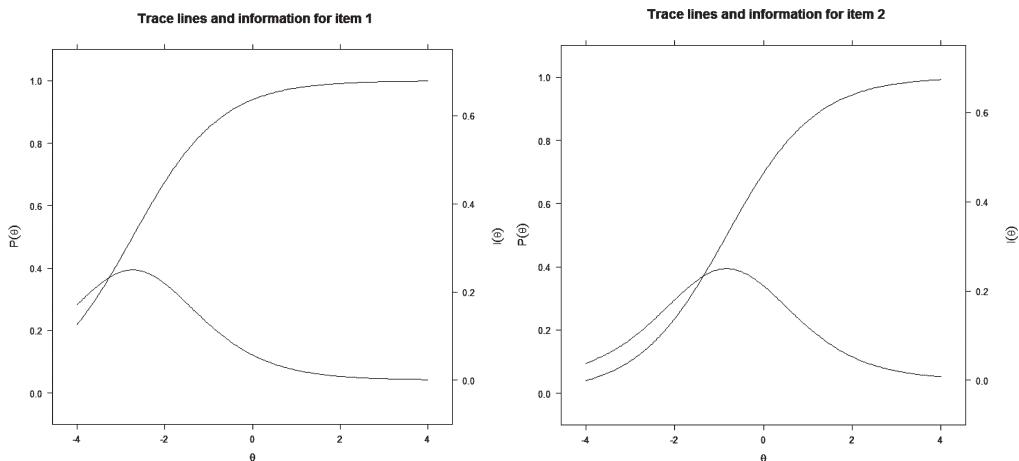


FIGURE 4.7. IRFs and item information functions for items 1 and 2.

Summary

In this chapter we presented a second estimation method, marginal maximum likelihood estimation, for instrument calibration. This method separates the estimation of items from that of persons by assuming that persons are randomly sampled from some population. In this approach, the original likelihood is multiplied by the population distribution and thereby eliminates, through integration, the person location parameters. MMLE then obtains MLEs of the item parameters by maximizing the (marginal) likelihood function. The numerical integration is accomplished by using a Gaussian quadrature approach. In short, the continuous population distribution is approximated by using a discrete distribution consisting of quadrature nodes and associated weights that reflect the density of the function around the node. As is the case with JMLE, MMLE is an iterative procedure. Once convergence is achieved, we assess model–data fit.

Because we separate item estimation from person estimation, we can use MLE or a Bayesian procedure, such as Bayes Modal Estimation or Bayes Mean Estimate (i.e., EAP), for estimating person locations. EAP is noniterative, and unlike MLE $\hat{\theta}$ s, EAP location estimates are available for all response patterns, including zero and perfect score patterns.

In addition to the model–data fit methods introduced in previous chapters, in this chapter we introduced a method for determining item parameter invariance by using confidence intervals and we covered a graphical fit approach using empirical and predicted IRFs. Specifically and with respect to the use of confidence intervals, if the CIs for corresponding items from two subsamples overlap, then we have evidence of invariance. With respect to the graphical fit approach, if the predicted IRFs show close agreement with the observed IRFs, then we have evidence supporting model–data fit. Moreover, the tenability of our model’s functional form assumption may be determined by examining the empirical IRFs. At this point, we have looked at model–data fit by using fit statistics

and graphically, both by determining the tenability of the dimensionality and functional form assumptions and by examining the invariance of item parameter estimates.

In this chapter, we introduced the transformation of one metric to another as well as the alignment or linking of two metrics. Metric transformation can be used to enhance the interpretability of our results. For instance, we can transform the person location estimate through the instrument's total characteristic function to a total score metric that respondents may find inherently informative. The total characteristic function relates the θ continuum to the expected trait scores on the instrument. In terms of metric alignment, we presented one approach for obtaining the metric transformation coefficients.

In the next chapter, we relax the constraint that all items must have a common discrimination parameter. Eliminating this constraint yields the two-parameter model. With this model there is a parameter for the item's location and another for its discrimination capacity. Allowing the discrimination parameter to vary across items allows the modeling of items that differentially discriminate. Therefore, the number of situations in which we may obtain model–data fit is potentially greater than what is possible with the one-parameter model.

Notes

1. Another alternative to using JMLE or marginal maximum likelihood estimation is to use Markov chain Monte Carlo (MCMC) simulation methods. The gist of MCMC involves conducting a simulation in which one randomly samples from a particular distribution. The MCMC method uses a subset of the generated data for the estimation of the parameters. Estimation of simple as well as complex models is comparatively easier with MCMC than with either JMLE or marginal maximum likelihood estimation because it does not require precalculation of derivatives (Patz & Junker, 1999a). However, MCMC implementation software is not as user friendly as the software available for JMLE and MMLE. Given this book's orientation, MCMC methods are not covered. Patz and Junker (1999a, 1999b) contain introductions to using MCMC methods with IRT models, and Baker and Kim (2004) discuss a particular MCMC method known as the Gibbs Sampler.
2. The symbol $\int_a^b f(x)dx$ means integrate the function $f(x)$ between a and b with respect to x ; dx means perform the integration with respect to x . The a and b are the limits of the integration, and integrals are sometimes called antiderivatives. For example, the integral $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-(z^2)/2} dz$ means integrate the expression $\frac{1}{\sqrt{2\pi}} e^{-(z^2)/2}$ over the range of values from a to b with respect to z . In integrating a function, one is finding the area under the function between a and b .
3. Implied in Figure 4.1 is that the quadrature points have special locations to maximize the accuracy of the approximation. As a contrarian example, one would not expect that using 11 quadrature points located between 0 and 1 would lead to an accurate approximation of the area under the curve in Figure 4.1. The locations (X_s) and their corresponding weights ($A(X_s)$) may be obtained from tables provided by Stroud

and Secrest (1966) for approximating the Gaussian error function. The Stroud and Secrest Gauss–Hermite X_r s and $A(X_r)$ s must be multiplied by $\sqrt{2}$ and $(1/\sqrt{\pi})$, respectively, to place them on the normal function metric (Bock & Lieberman, 1970). However, typically the Stroud and Secrest values are not used in programs that use MMLE, such as BILOG-MG. Rather, a specified range of the θ continuum (e.g., -4.0 to 4.0) is divided into R equidistant discrete points that serve as the X_r s, and the standard unit normal probability density is computed at each of the R points (i.e., $f(X_r) = (\frac{1}{\sqrt{2\pi}})e^{-(X_r^2)/2}$). The probability density at X_r is multiplied by the interval width (i.e., $X_r - X_{r+1}$) to obtain the quadrature weight $A(X_r)$. If the discrete (prior) distribution is symmetric, then the $A(X_r)$ s need to be calculated only for the X_r s ≤ 0 . The X_r s and $A(X_r)$ s must satisfy the constraints $\sum A(X_r) = 0.0$, $\sum X_r A(X_r) = 0.0$, and $X_r \sum X_r^2 A(X_r) = 1.0$.

4. An alternative to using biweights or a Bayesian strategy is to use a *weighted MLE* approach for estimating θ (Warm, 1989) or an *AMT-Robustified Jackknife MLE* (Wainer & Wright, 1980). The former approach uses, in effect, the instrument's total information to reduce the bias inherent in MLE $\hat{\theta}$ s. The latter approach uses the Jackknife procedure to accomplish the same thing when the bias is due to response aberrations. Wainer and Wright (1980) investigated the AMT-Robustified Jackknife MLE and found that it performed well in the context of guessing; the study assumed that reasonably good estimates of δ_j could be obtained, for example, by culling from the calibration sample persons with unusual response patterns. Also see Weitzman (1996).
5. Specifying a Rasch calibration with BILOG-MG version 3.0 requires, in addition to specifying the 1PL for Model in the Setup menu's General . . . dialog, that one also select One Parameter Logistic Model from the Technical menu's Calibration Options . . . dialog. (Some would consider the One Parameter Logistic Model check box to be mislabeled.) Selecting this check box instructs the program to rescale the 1PL estimates so that $\alpha = 1$; the δ_j s are also appropriately transformed. How this rescaling is performed is discussed in the "Metric Transformation and the Total Characteristic Function" section.
6. Although the presence of omitted item(s) at the end of an instrument may indicate speededness, their absence does not necessarily indicate that individuals had sufficient time to take the instrument. This is because an individual realizing that administration time is about to expire may simply answer at random those items he or she has insufficient time to appropriately consider.
7. The biserial correlation coefficient (ρ_b) is a measure of the association between an artificially dichotomized variable and a continuous variable. With a ρ_b one has two normally distributed continuous variables, but for some reason one variable is transformed to two categories. The nondichotomized continuous variable may be the variable being measured by the instrument. In contrast, the point-biserial (ρ_{pb}) is an association index for a genuine dichotomous variable and a continuous variable; the point-biserial is a special case of the Pearson correlation coefficient. The con-

tinuous variable is assumed to be normally distributed. The relationship between these correlations is presented in Appendix C.

8. The term *threshold* should not be confused with the epidemiological use of *threshold* in logistic regression. In that context, a threshold is the point on a continuum where the response function first begins to increase rapidly. Therefore, the associated probability with this threshold is less than 0.5. In BILOG, *threshold* refers to the item's location.
9. Some versions of BILOG, for example, BILOG-3, also contain a column labeled DISPERSION. DISPERSION is the item standard deviation when working with the normal ogive models and is equaled to $1/\alpha$; a normal ogive model is discussed in Appendix C.
10. Although the estimation algorithm is searching for a maximum, a function may also have a minimum (or minima). The convergence criterion can be satisfied when the algorithm has found a maximum or a minimum. Therefore, the Second-order test determines whether we have a maximum or a minimum. The message Second-order test: model is a possible local maximum examines the information matrix to determine if it is positive definite (i.e., has positive eigenvalues). If this matrix of partial second derivatives (a.k.a., Hessian matrix) is positive definite then we have a maximum. In our case, the estimation algorithm has found a maximum that is not on a boundary, and we have a “sensible” solution. As a result, we can proceed to interpret our results.
11. Although the program seeks to create equal-size groups, this does not mean that the variability of ability for a group is constant across groups. For instance, for the lowest fractile (`theta = -1.6434`) the program will begin with the least able person and accumulate successively more able people until reaching 3268 examinees ($\sim N/group.bins = 19,601/6$); a large percentage of these least able individuals will respond incorrectly to the non-extreme item of interest. Because the lowest fractile is located where there are comparatively few individuals, the range of ability in this fractile will be greater than for a fractile where many persons are located such as in the center of the distribution (e.g., fractile 4). At the upper end of the continuum, there will be a tendency to see more correct responses than expected because the accumulation at the upper end will tend to group people that respond correctly to a non-extreme item.
12. Figure 4.8 contains an example figure using the `itemfit` argument `empirical.CI = .95` with a data set composed of 1000 cases. Ideal `itemfit` would have the response function within each error bar.
13. Calling the `itemfit` function with the `empirical.table` argument (i.e., `itemfit(rasch, empirical.table = 3, group.bins = 6)`) allows us to see the specific theta values used in Figure 4.6 as well as the corresponding expected and observed frequencies of our responses. For example, for our lowest group (`theta = -1.6434`), we expected to see 1001.106 correct responses, but we only observed 213; for the second fractile, the discrepancy between what we would

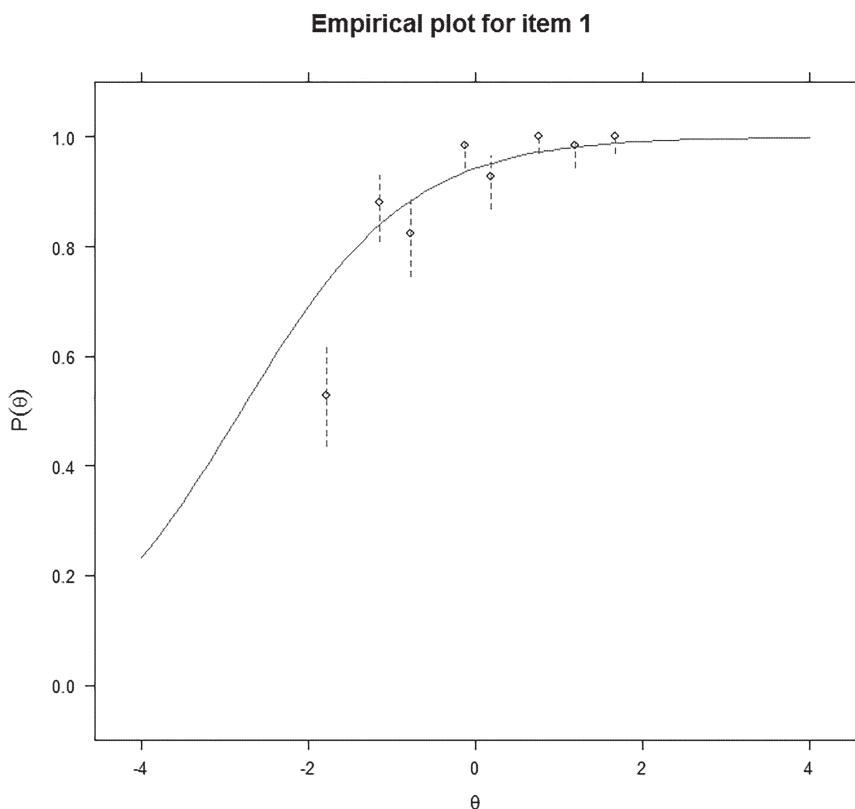


FIGURE 4.8. IRF with observed proportions and error bars.

predict and what we observed is only 112.337 (or 0.5731% of 19,601 examinees); for our third fractile, the observed number of correct responses is very close to what we would predict (Observed = 2052, Expect = 2048.15); and so on. This series of tables is useful for diagnosing misfit with the itemfit plot and/or specific fit statistics for diagnostic purposes (e.g., McKinley & Mills' [1985] G^2). The sum of a fractile's cat _ 0 and cat _ 1 observed frequencies is the number of examinees in the fractile.

```
> itemfit(rasch,empirical.table=2,group.bins=6)
$`theta = -1.6434`           ← Fractile 1
    Observed Expected z.Residual
cat _ 0      3055  2266.894  16.55272
cat _ 1       213   1001.106 -24.90834

$`theta = -0.8183`           ← Fractile 2
    Observed Expected z.Residual
cat _ 0      1739  1626.663  2.785311
cat _ 1      1527  1639.337 -2.774523

$`theta = -0.3062`           ← Fractile 3
    Observed Expected z.Residual
```

```
cat_0      1214  1217.85 -0.11032653
cat_1      2052  2048.15  0.08507384

$`theta = 0.2439`           ← Fractile 4
    Observed Expected z.Residual
cat_0      826   834.128 -0.2814291
cat_1     2440  2431.872  0.1648219

$`theta = 0.832`           ← Fractile 5
    Observed Expected z.Residual
cat_0      143   522.6174 -16.605585
cat_1     3123  2743.3826  7.247746

$`theta = 1.6914`           ← Fractile 6, the  $\hat{\theta}$  when  $X = 5$ 
    Observed Expected z.Residual
cat_0      0    244.0094 -15.620802
cat_1     3269  3024.9906  4.436542
```

14. To calculate the correlation, M_s and SD_s , for each of our calibration's $\hat{\delta}_s$, we extract the estimates from each of the output objects, `rasch1` and `rasch2`, by using the `extract.item` function. This function returns intercept values (γ_j) rather than item locations (δ_j). Because we are working with the Rasch model, these γ_j s are easiness values. That is, given Equation 2.3 and $\alpha = 1.0$, we have $\gamma_j = -\alpha\delta_j = -\delta_j = \delta_j^E$. Below we convert these easiness values to a difficulty scale to be consistent with the use of $\hat{\delta}_s$ in Table 4.6. The code for one way of doing this is as follows.

```
> nitems = 5
> # subsample 1
> s1_itest=matrix(data=0,nrow= nitems,ncol=4)      # create a null matrix to
hold the item est

> # subsample 2
> s2_itest=matrix(data=0,nrow= nitems,ncol=4)

> # one way to extract item est from subsample 1 & 2
> for(i in 1: nitems){
  s1_itest[i,]=extract.item(rasch1,i)@par
  s2_itest[i,]=extract.item(rasch2,i)@par }

> s1_itest=as.data.frame(s1_itest)                  # subsample 1: convert matrix to
data frame
> s2_itest=as.data.frame(s2_itest)                  # subsample 2: convert matrix to
data frame

> colnames(s1_itest)=c('a','b','g','u')          # subsample 1: meaningful variable
names
> s1_itest                                         # 'a' is alpha, 'b' is delta
   a      b      g      u
1 1  2.7462109 -999 999
2 1  0.8369125 -999 999
3 1  0.3799882 -999 999
4 1 -0.3813749 -999 999
5 1 -0.6451739 -999 999
```

```

> colnames(s2_itest)=c('a','b','g','u')                                # subsample 2: meaningful variable
names
# 'a' is alpha, 'b' is delta

> s2_itest
      a      b      g      u
1 1  2.7268362 -999 999
2 1  0.8153803 -999 999
3 1  0.3704098 -999 999
4 1 -0.4236428 -999 999
5 1 -0.6154304 -999 999

> # convert easiness to difficulty
> s1_itest$b=s1_itest$b*-1
> s2_itest$b=s2_itest$b*-1

> s1_itest
      a      b      g      u
1 1 -2.7462109 -999 999
2 1 -0.8369125 -999 999
3 1 -0.3799882 -999 999
4 1  0.3813749 -999 999
5 1  0.6451739 -999 999

> s2_itest
      a      b      g      u
1 1 -2.7268362 -999 999
2 1 -0.8153803 -999 999
3 1 -0.3704098 -999 999
4 1  0.4236428 -999 999
5 1  0.6154304 -999 999

> print((s1_bmean=mean(s1_itest$b))) # subsample 1
[1] -0.5873126

> print((s2_bmean=mean(s2_itest$b))) # subsample 2
[1] -0.5747106

> # one way to transform subsample 2's metric to subsample 1's:
# mean-mean (see Ch 11)
> kappa=(s1_bmean - s2_bmean)      # Equation 4.21 because zeta = 1 with the
Rasch model
> kappa
[1] -0.01260195

> # only item locations transformed
> # s2_itest$b=s2_itest$b + kappa          # Eq 4.17; if only inter-
ested in transforming locations

```

15. To align our CI metrics, we use our kappa from Endnote 14. We extract subsample 2's item parameter estimates along with their CIs and store them in `s2_itest`. Subsequently, we apply Equation 4.17 to the elements of `s2_itest` to produce the aligned metric values saved to `s2_iteststar`. We present two alternative approaches. The first, the minimalist approach, simply transforms the extracted subsample 2 estimates, including discrimination (a), guessing (g), and the upper

asymptote (u). Consequently, one needs to ignore a , g , and u when comparing the subsample 2 CIs to subsample 1s (i.e., `print((s1_itest = coef(rasch1, IRTpars = TRUE)))`). The second method extracts the subsample 2 estimates and manipulates them for ease of comparison.

```
> s2_itest=coef(rasch2,IRTpars=TRUE)           # subsample 2: extract item
                                                parameters/CI

> # method 1; ignore a, g & u which have also been adjusted
> s2_iteststar=lapply(s2_itest, function(x) (x + kappa))          # apply Equation 4.17
> s2_iteststar
$II1
      a        b        g        u
par   0.9873981 -2.739438 -0.01260195 0.9873981
CI _ 2.5       NA -2.791459       NA       NA
CI _ 97.5      NA -2.687417       NA       NA

$II2
      a        b        g        u
par   0.9873981 -0.8279822 -0.01260195 0.9873981
CI _ 2.5       NA -0.8740566       NA       NA
CI _ 97.5      NA -0.7819078       NA       NA
:

> # method 2
> s2_mtrx=matrix(0,nrow =nitems, ncol=3)
> for (i in 1:nitems){
+ for (j in 1:3) s2_mtrx[i,j]=s2_itest[[i]][j,2]}          # metrics not aligned
> s2_mtrx
      [,1]      [,2]      [,3]
[1,] -2.7268362 -2.7788572 -2.6748152
[2,] -0.8153803 -0.8614547 -0.7693059
[3,] -0.3704098 -0.4189234 -0.3218962
[4,]  0.4236428  0.3666284  0.4806573
[5,]  0.6154304  0.5556629  0.6751979

> # transform subsample 2's item locations & CIs to subsample 1's
  metric
> s2_iteststar=apply(s2_mtrx, 2,function(x) (x + kappa))          # apply Equation 4.17
> s2_iteststar                                         # metric aligned to subsample 1
      [,1]      [,2]      [,3]
[1,] -2.7394381 -2.7914591 -2.6874171
[2,] -0.8279822 -0.8740567 -0.7819078
[3,] -0.3830117 -0.4315253 -0.3344982
[4,]  0.4110409  0.3540264  0.4680553
[5,]  0.6028284  0.5430610  0.6625959

> s2_iteststar=as.data.frame(s2_iteststar)
> names(s2_iteststar) = c('loc','CI _ 2.5','CI _ 97.5')          # meaningful
  variable names
> s2_iteststar
```

```

loc      CI _2.5    CI _97.5
1 -2.7394381 -2.7914591 -2.6874171
2 -0.8279822 -0.8740567 -0.7819078
3 -0.3830117 -0.4315253 -0.3344982
4  0.4110409  0.3540264  0.4680553
5  0.6028284  0.5430610  0.6625959

> # for ease of comparison convert subsample 1's estimates to matrix
   format
> s1_itest=coef(rasch1,IRTpars=TRUE)
> s1_mtrx=matrix(0,nrow =nitems, ncol=3)
> for (i in 1:nitems){
+   for (j in 1:3) s1_mtrx[i,j]=s1_itest[[i]][j,2]}
> s1_mtrx=as.data.frame(s1_mtrx)
> names(s1_mtrx) = c('loc','CI _2.5','CI _97.5')

```

See Chapter 11 for more information on the alignment of metrics.

16. To perform a 1PL calibration with mirt, one has to impose constraints on the item discriminations to be equal using mirt model. The corresponding output object is then passed to mirt:

```

> OnePLconstr=mirt.model('Theta=1-5
+ CONSTRAIN = (1-5,a1)')
> OnePL=mirt(mathdata,OnePLconstr,SE=T,SE.type='Fisher')

```

Our estimates are:

```

> print(coef(OnePL,IRTpars=TRUE,printSE=T),digits=5)
$I1
      a      b      g      u
par 1.44098 -1.89948  0    1
SE  0.01728  0.02574 NA NA

$I2
      a      b      g      u
par 1.44098 -0.57346  0    1
SE  0.01728  0.01433 NA NA

$I3
      a      b      g      u
par 1.44098 -0.26045  0    1
SE  0.01728  0.01318 NA NA

$I4
      a      b      g      u
par 1.44098  0.27941  0    1
SE  0.01728  0.01310 NA NA

$I5
      a      b      g      u
par 1.44098  0.43747  0    1
SE  0.01728  0.01354 NA NA

$GroupPars

```

```
MEAN _ 1 COV _ 11
par      0      1
SE       NA     NA
```

17. Figure 4.9 shows the TCF using the `mirt` `plot` function:
`plot(rasch,type = "score", theta_lim = c(-4,4)).`

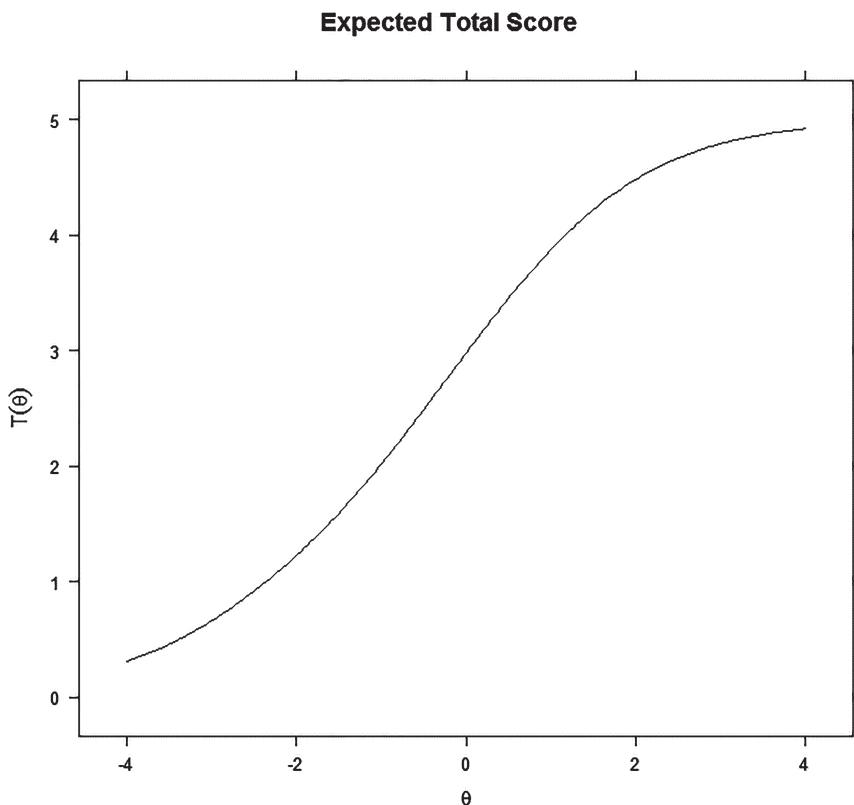


FIGURE 4.9. Total characteristic curve using `mirt`.

5

The Two-Parameter Model

In Chapter 2 we developed a model based on the premise that the distance between a person's location and an item's location is an important determinant of his or her response. However, when we examine a traditional discrimination index, such as the biserial correlation, we typically find that it varies across the items on the instrument. For instance, the biserials for the mathematics data vary from 0.397 to 0.564 (see Table 4.3). In this chapter we extend our distance idea to incorporate information about how well an item discriminates among individuals located at different points. We still use the distance between the person's and the item's locations, but we now weight this distance by how well the item discriminates. Thus, the probability of a response of 1 is a function of not only how far apart the person and the item are, but also how well the item differentiates among respondents located at different points on the continuum.

By taking into account how well an item discriminates, we are relaxing the constraint that items must share a common slope. As a result, we are able to obtain model fit in a greater number of situations than with the Rasch model. Our relaxing of the common discrimination parameter constraint implies a philosophical shift. With the Rasch model our interest is not so much in modeling the data, but rather in constructing an instrument that is consistent with the Rasch model.¹ In contrast, with the two-parameter model our philosophical perspective is one of modeling the data. In the following discussion, we conceptually develop the two-parameter logistic model. Consistent with the previous three chapters, we then apply the model to the mathematics data introduced in Chapter 2. As part of our model fit analysis, we revisit item parameter invariance and discuss strategies for its examination when item discrimination is allowed to vary across items. We end by discussing relative efficiency as an approach for simultaneously comparing multiple total information functions.

Conceptual Development of the Two-Parameter Model

As is true for the 1PL model, the data for the two-parameter model need to be dichotomous. Such data may come from, for example, a personality inventory, a depression

scale, or an examination (cf. Reise & Waller, 1990; Schaeffer, 1988). In Chapter 2 we pointed out we could potentially improve fit by varying the slope of the predicted IRF to more closely match that of the empirical IRF.

Figure 5.1 contains the IRFs for five items with different discriminations but located at the same point on the continuum. This common location corresponds to the intersection point of all five IRFs and is $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 1.0$. As can be seen, as the values of α change from 0.5 to 3.0, the corresponding IRFs become progressively steeper. For example, comparing item 1 ($\alpha_1 = 0.5$) with item 5 ($\alpha_5 = 3.0$), one sees that item 1's slope is substantially less than that of item 5. These IRFs may be modeled by modifying the 1PL model to allow for α to vary across items. When done, the 1PL model becomes the Birnbaum two-parameter logistic (2PL) model

$$p(x_j = 1 | \theta, \alpha_j, \delta_j) = \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} = \frac{e^{\alpha_j\theta + \gamma_j}}{1 + e^{\alpha_j\theta + \gamma_j}}, \quad (5.1)$$

where θ is the person location parameter, and δ_j and α_j are item j 's location and discrimination parameters, respectively, and the intercept (constant) is $\gamma_j = -\alpha_j\delta_j$; the subscript on α indicates that each item j has its own discrimination parameter. With the 2PL model, the logistic deviate or logit, $\alpha_j(\theta - \delta_j)$, contains the item's two parameters, δ_j and α_j .² (The 2PL model may also be written to include the scaling constant, D ; see Appendix C.) The 2PL model is predicated on a unidimensional latent space, condi-

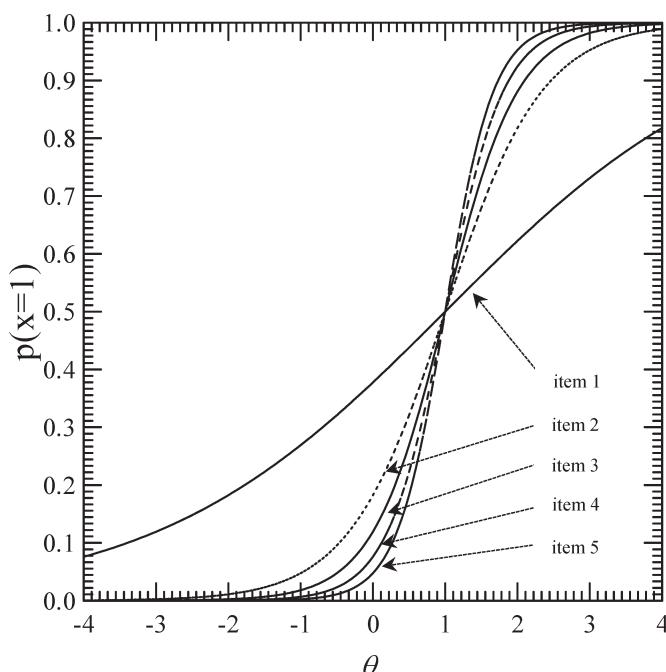


FIGURE 5.1. IRFs for five items with the same location and different slopes ($\alpha_1 = 0.5, \alpha_2 = 1.5, \alpha_3 = 2.0, \alpha_4 = 2.5, \alpha_5 = 3.0; \delta = 1.0$ for all five items).

tional independence, and the functional form assumptions discussed in Chapter 2. The functional form assumption embodies an IRF with a lower asymptote of 0. For ease of presentation we use p_j instead of $p(x_j = 1 | \theta, \alpha_j, \delta_j)$ in the following discussion.

An item's α_j characterizes how well the item can differentiate among individuals located at different points on the continuum. As is the case with the 1PL model, α_j is proportional to the slope of the IRF at its inflection point, δ_j . The slope at δ_j is $0.25 * \alpha_j$.³ As the value of α_j increases, the IRF's slope becomes steeper and the item's capacity to discriminate between individuals increases. When items vary in their discrimination, then their corresponding IRFs cross one another at some point along the continuum.

The discrimination parameter can theoretically vary from $-\infty$ to ∞ , with an $\alpha = \infty$ (or $\alpha = -\infty$) reflected in a step function IRF. Reasonably "good" values of α range from approximately 0.8 to about 2.5. A negative α_j reflects an item where individuals with lower θ s have a higher probability of obtaining a response of 1 than individuals with higher θ s (i.e., a monotonically nonincreasing IRF). As such, an item with a negative α_j is behaving in a counterintuitive fashion. Similar to a traditional negative discrimination index (e.g., a negative point biserial or biserial correlation), a negative α_j may indicate an item that should be discarded because its performance is inconsistent with the model or, in the case of proficiency assessment, an item that has its correct response incorrectly specified.

Figure 5.1 shows that a 2PL model's IRFs share a similarity with those of the 1PL model. Namely, the item is located at the point where an individual randomly selected from all the persons located at the item's location has a 50–50 chance of obtaining a response of 1. Because this is also the point of maximum slope, the item location is the point at which the item discriminates most effectively among respondents.

Information for the Two-Parameter Model

There is a relationship between an item's α_j and how much it reduces our uncertainty about where a person is located on the continuum. As an item's discrimination parameter increases, the maximum item information for estimating θ increases, and this in turn leads to a decrease in our uncertainty about a person's location.⁴ This increase in available information at δ_j is associated with a concomitant decrease in the standard error for our estimated person location at this point. Recall that the standard error is our measure of uncertainty. Because this asymptotic standard error reflects the variability of the $\hat{\theta}$ s over an infinite number of independent administrations, a reduction in the standard error means the $\hat{\theta}$ s would be tightly clustered together about θ and we would have greater certainty about the person's true location.

The item information functions ($I_j(\theta)$ s) corresponding to the items in Figure 5.1 are presented in Figure 5.2. As is the case with the 1PL model, Figure 5.2 shows that with the 2PL an item provides its maximum information at δ_j and that the distribution of information is unimodal and symmetric about δ_j . However, we see that for individuals located at or near $\delta = 1.0$, these items vary in their amount of information for estimating

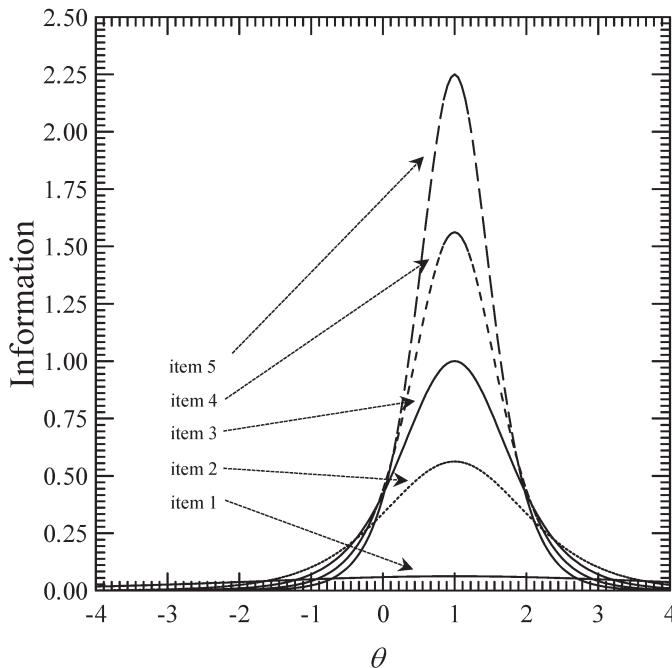


FIGURE 5.2. Item information functions for five items from Figure 5.1 ($\alpha_1 = 0.5$, $\alpha_2 = 1.5$, $\alpha_3 = 2.0$, $\alpha_4 = 2.5$, $\alpha_5 = 3.0$; $\delta = 1.0$).

a person's location. In contrast to the 1PL model, the maximum amount of information provided by an item varies as a direct function of the magnitude of α_j . Item 5 with an $\alpha_5 = 3.0$ provides the greatest amount of information for distinguishing among individuals in the vicinity of δ_j , whereas item 1 with the smallest α ($\alpha_1 = 0.5$) provides the least information at δ_j .⁵

In Chapter 2 a general formulation of item information is presented. Specifically, item information is given by

$$I_j(\theta) = \frac{[p'_j]^2}{p_j(1-p_j)}. \quad (5.2)$$

Conceptually, Equation 5.2 can be interpreted as an item's information at θ equals how quickly the IRF changes over the (conditional) variance at θ .⁶ Because the 2PL model's first derivative is

$$p'_j = \alpha_j p_j (1-p_j), \quad (5.3)$$

its substitution for p'_j in Equation 5.2 produces the 2PL model item information function

$$I_j(\theta) = \alpha_j^2 p_j (1-p_j). \quad (5.4)$$

Because when $p_j = (1 - p_j) = 0.5$ the product $p_j = (1 - p_j)$ is at its maximum, our maximum item information is $0.25\alpha_j^2$. When an item is calibrated with the two-parameter model and $\alpha_j > 1.0$, then the item contributes more information for estimating θ than when the Rasch model is used for calibration. Moreover, we see that the 1PL model's item information, $\alpha_j^2 p_j = (1 - p_j)$, is a special case of Equation 5.4. The total information for an instrument is defined as it is in Chapter 2 (i.e., the sum of the item information functions)⁷

$$I(\theta) = \frac{1}{\sigma_e^2(\theta)} = \sum_{j=1}^L I_j(\theta). \quad (5.5)$$

Conceptual Parameter Estimation for the 2PL Model

If we were to graphically present the log likelihood function for estimating δ_j in the Rasch model, we would need only two dimensions because item discrimination is constant. That is, one axis represents δ , whereas the other reflects the log likelihood values. The resulting figure would look similar to the $\ln L$ shown in Figure A.1 (in Appendix A). However, when estimating an item's α_j and δ_j the graphical presentation of the (log) likelihood function requires three dimensions (Figure 5.3), with one dimension for each parameter and the third representing the (log) likelihood values. (This is analogous to having a regression line with one predictor [i.e., the Rasch model case], but requiring a regression plane with two predictors [i.e., the 2PL model case].) We can view the log likelihood curve for estimating δ_j in the Rasch model as a “slice” through the surface (Figure 5.3) conditional on α .

As we see, as α_j increases, the surface becomes more peaked and, conversely, as α_j decreases, the surface becomes less peaked. Therefore, as α_j decreases, it becomes more difficult to accurately estimate δ_j and, conversely, as α_j increases, the estimation of δ_j becomes comparatively easier, all other things being equal. Our estimation requires us to simultaneously determine the location of the maximum of this surface with respect to both the α and δ axes. Conceptually, one has a plane that is tangent to the surface. This plane moves along the surface until it arrives at the maximum of the log likelihood surface. At this point, the plane's slope with respect to both α and δ are both zero (i.e., the plane is horizontal). The location of the maximum on the α axis is our $\hat{\alpha}$, and the location on the δ axis is our $\hat{\delta}$. In Figure 5.3 the maximum of the $\ln L$ occurs at approximately an $\hat{\alpha}$ of 2.2 and $\hat{\delta}$ of -1.6. The log likelihood surface shown is item 1's observed log likelihood surface.⁸

The 2PL model's log likelihood equation is seen in Chapter 2 (Equation 2.10) but with p_j determined by Equation 5.1. The general principles for parameter estimation that were outlined in Chapters 2 and 3 (as well as Appendices A and B) also apply here. Of course, the first and second derivatives are different from those of the Rasch/1PL model. The mathematical details for 2PL model estimation are shown in Baker and Kim (2004). The 2PL model's parameters may be estimated by different methods such as JMLE or MMLE. As is the case with the 1PL model, starting values for the estimation need to be provided. Equations C.12 and C.16 (Appendix C) can be used to provide these starting values for estimating α_j and δ_j , respectively.

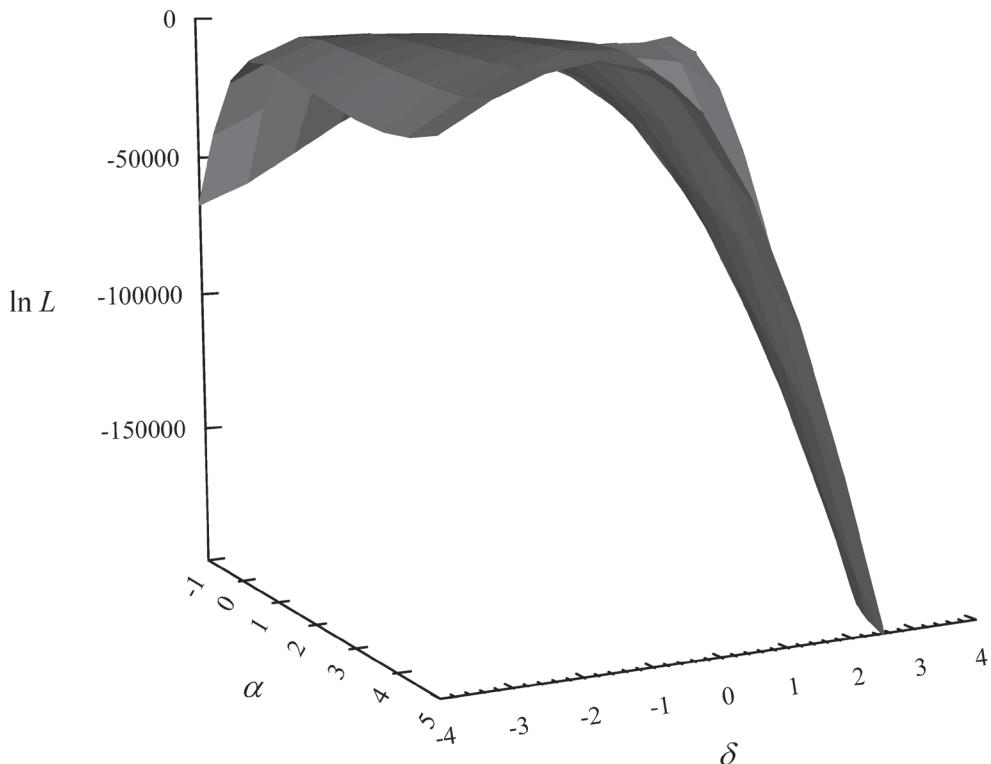


FIGURE 5.3. Observed log likelihood surface for item 1; $\hat{\alpha} \approx 2.2$ and $\hat{\delta} \approx -1.6$.

The 2PL model, like the 1PL model, has sufficient statistics. For estimating a person i 's location the sufficient statistic is the weighted sum of the item responses, $\sum \alpha_j x_{ij}$.⁹ In general, this weighted sum provides more information than the unweighted sum (i.e., $\sum x_{ij}$) except when the two sums are identical or proportional (Lord, 1983a). One implication of weighting the item responses by the item discrimination parameters is that different patterns of responses, albeit with the same observed score (e.g., 11100 and 10011), result in different person location estimates. However, unlike the sufficient statistic for the 1PL model, the sufficient statistic for person estimation in the 2PL model depends on the unknown discrimination parameters, α_j s. In short, the sufficient statistic for estimating a person's location is not independent of the items because we need to know their α_j s. If the estimates of the α_j s are inaccurate, then the weighted sum is less informative than the unweighted sum (Lord, 1983a). The 2PL model has a sufficient statistic, $\sum p_j \theta_i$, for estimating α_j (Baker & Kim, 2004). However, this sufficient statistic also requires knowledge of a parameter—the person parameter.

How Large a Calibration Sample?

A number of studies have investigated the accuracy of parameter estimation under various conditions, such as with different instrument lengths, sample sizes, and estimation

methods. For instance, in a study of MMLE, Drasgow (1989) found that as few as 200 persons and five items were required for “essentially” unbiased parameter estimates with reasonably small standard errors as long as α and δ were not too extreme. Seong (1990b), using a 45-item instrument, studied how the accuracy of MMLE parameter estimation was affected by θ distribution shape (normal, positively skewed, negatively skewed), the number of quadrature points (10 or 20), the sample size ($N = 100$ or 1000), and the assumed prior θ distribution (normal, positively skewed, negatively skewed); BILOG was the calibration program. Overall, the estimated item location and discrimination parameters were more accurate when the prior distribution matched the θ distribution than when there was a mismatch; the sample size was 1000. Moreover, under the match θ /prior distributions and $N = 1000$ condition, the accuracy of the item parameter estimates could be improved by increasing the number of quadrature points. Seong suggested that the default normal prior distribution and quadrature points used in BILOG were reasonable choices for a small sample size.

In another Monte Carlo study, Stone (1992) examined the effect of instrument length ($L = 10, 20$, or 40), sample size ($N = 250, 500$, or 1000), and θ distribution shape (normal, positively skewed, or symmetric and platykurtic) on estimation accuracy with a MMLE program, MULTILOG (Thissen, Chen, & Bock, 2003). In general, he found that with 500 or more individuals and instruments of 20 or more items, both item location and discrimination estimates were generally precise and stable. The effect of the θ distribution factor was ameliorated by using the 40-item instrument. In contrast, Harwell and Janosky (1991) looked at estimation accuracy in the context of prior distribution characteristics with BILOG. They found that for short instruments (15 items) and small samples (e.g., $N = 75, 100, 150$) the prior variances affected the accuracy of parameter estimation. Comparatively longer instruments (25 items) were not affected as much by the prior distribution variance when the sample sizes were greater than 100. They recommended that for 15-item instruments and fewer than 250 persons a prior variance of 0.25 not be used for α ; all simulees were randomly sampled from a normal distribution and restricted to the range -3 to 3.

Strictly speaking, because these are fixed effects design studies we should not generalize beyond the conditions that were investigated. However, these studies may still provide some guidance as to what one might anticipate to occur in somewhat similar situations. For example, the results indicate that a number of factors (e.g., instrument length, prior distributions, estimation method) affect item parameter estimation accuracy and thus need to be considered in determining calibration sample size. Additional factors to take into account in determining a desirable sample size target is the application’s purpose, ancillary technique sample size requirements, model–data fit tolerance, the estimation method, instrument characteristics, the variability and distribution of respondents, and the amount of missing data. The interaction of these variables makes it difficult to provide a suggested sample size that would be applicable in all situations. However, to provide some guidance, we offer an admittedly rough guideline.

Assuming MMLE, the use of a prior distribution for α , and favorable conditions (e.g., θ /prior distribution match, etc.), it appears that a calibration sample size of at least 500 persons and instruments of 20 or more items tend to produce reasonably accurate

item parameter estimates. It must be noted that less favorable situations may necessitate larger sample sizes. However, it should also be observed that “it is not necessarily true that because the parameters of a fitted function are not well estimated that certain other characteristics of the function are unstable . . . like the information function” (Thissen & Wainer, 1982, p. 409). That is, there may be situations where even though item parameters may not be well estimated, the estimates may still be useful. Moreover, it may be anticipated that there is a sample size, say 1,200, at which one reaches, practically speaking, a point of diminishing returns in terms of improvement in estimation accuracy, all other things being equal; this should not be interpreted as an upper bound. To reiterate, the caveats and considerations previously mentioned, as well as the need to avoid interpreting sample size guidelines as hard-and-fast rules, still apply to our recommendation.¹⁰

Metric Transformation, 2PL Model

Examination of the model in Equation 5.1 shows that the indeterminacy issue raised with the 1PL model also applies to the 2PL model. Namely, we can add or subtract a constant from θ and δ_j and not change the logistic deviate. As a result, the IRF is unaffected, although its location moves up or down the continuum. Stated another way, the origin of the metric is arbitrary. Similarly, multiplying θ and δ_j by a constant and dividing α_j by the same constant would leave $\alpha_j(\theta - \delta_j)$ unchanged. This implies that the unit for measuring θ and δ_j is also arbitrary. As discussed in Chapters 3 and 4, this indeterminacy is addressed in different ways such as through person centering.

This metric indeterminacy facilitates the transformation of the metric to have desired characteristics. As discussed in Chapter 4, we can rescale our parameters or their estimates by using our metric transformation coefficient ζ and κ . The discrimination parameter (or its estimate) for each item is transformed by

$$\alpha_j^* = \frac{\alpha_j}{\zeta}, \quad (5.6)$$

where α_j is item j 's discrimination on the metric to be transformed (i.e., α_j is on the initial metric) and α_j^* is the transformed discrimination value (i.e., α_j^* is on the target metric). In terms of a slope–intercept parameterization, the item-wise transformation would be

$$\gamma_j^* = \gamma_j - \frac{\alpha_j(\kappa)}{\zeta}. \quad (5.7)$$

In general, our linear transformation is $\xi^* = \zeta(\xi) + \kappa$, where ζ and κ are the unit and location of the new metric, respectively. To transform our item locations, ξ represents δ_j (or its estimate) on the initial metric and ξ^* reflects δ_j^* (or its estimate) on the target metric

$$\delta_j^* = \zeta(\delta_j) + \kappa. \quad (5.8)$$

In addition, we can transform the person locations by letting ξ be θ (or its estimate) on the initial metric, with ξ^* representing the transformed person location θ^* (or its estimate). The values of ζ and κ may be given for a particular scale, such as the T-score scale, or they may be obtained by Equations 4.20 and 4.21. However, Equations 4.20 and 4.21 ignore information in the item discrimination parameters. Therefore, when item discrimination varies across items, a preferable method for determining the values of ζ and κ is the total characteristic function equating approach presented in Chapter 11.

As is the case with the 1PL model, with the 2PL model the instrument has a total characteristic curve that is bounded by 0 and the instrument's length, L. The nonlinear transformation of θ to the total score metric is accomplished, as shown in Chapter 4, by $T = \sum_{j=1}^L p_j$, where p_j is given by the 2PL model.

Example: Application of the 2PL Model to the Mathematics Data, MMLE, BILOG-MG

Various programs can be used to perform our calibration, such as flexMIRT, IRTPRO, the R packages TAM and mirt), Mplus (Muthén & Muthén, 2017), NOHARM, SAS proc irt, or XCALIBRE (Assessment Systems Corporation, 1997), to name just a few. (Using NOHARM for calibration is demonstrated in Appendix G, “Example: NOHARM Unidimensional Calibration.”) For our example we use BILOG-MG to perform the calibration of the mathematics data introduced in Chapter 2. Subsequently, we use mirt to perform a 2PL model calibration; in Chapter 10 flexMIRT is used for a 2PL model calibration.

In Chapter 3 we assess the tenability of the unidimensionality assumption. For brevity we do not repeat the analysis here.¹¹ The command file for our calibration (Table 5.1) is similar to the one used for the 1PL model (Table 4.2) except for a few modifications. These changes replace NPARM = 1 with NPARM = 2 to specify the two-parameter model (GLOBAL line) and obtain person parameter estimates as part of the item calibration; LOG indicates the logistic version of the two-parameter model (i.e., a 2PL model calibration). In this example, we perform person and item parameter estimation in a single run for two reasons. First, we want to show how to estimate the item and per-

TABLE 5.1. The BILOG Command File for the 2PL Model Item Calibration

```
2PL item calibration

>GLOBAL DFNAME='MATH.DAT', NPARM=2, NWGHT=0, LOG, SAVE;
>SAV PARM='MATH.PAR', SCO='MATH.SCO';
>LENGTH NITEMS=5;
>INPUT NTOT=5, NALT=2, NIDC=10, SAMP=20000, TYPE=1;
>ITEMS;
>TEST TNAMES='MATH',
    INumber = (1(1)5);
    (10A1,T1,5(1X,1A1))
>CALIB CYCLES=20, NEWTON=20, CHI=(5,9), PLOT=1.0;
```

son parameters in a single run. Second, because in Chapters 3 and 4 we concluded that we have acceptable model–data fit with the Rasch model, we expect the less restrictive 2PL model to also show acceptable model–data fit. Therefore, our rationale in Chapter 4 for separating person and item parameter estimation is not applicable.

To estimate person locations we include the SCORE command line. Although both MAP and MLE are available (with or without the biweight modification), we use EAP person parameter estimation, >SCORE MET = 2, NOPRINT.¹² By default, BILOG prints all the $\hat{\theta}$ s to the Phase 3 listing file. However, because we do not want 19,601 $\hat{\theta}$ s in our listing file we suppress this printing by using the subcommand NOPRINT on the SCORE line. Therefore, to see our $\hat{\theta}$ s we instruct BILOG to save them to an alternative file by using the SAV command line with the subcommand SCO = 'MATH.SCO' and the SAVE subcommand on the GLOBAL line. This alternative file, MATH.SCO, can also be used in further statistical analyses of the $\hat{\theta}$ s.

Table 5.2 contains the Phase 1 and 2 results. The line ITEM RESPONSE MODEL shows that the calibration is using the 2 PARAMETER LOGISTIC model. The Phase 2 output shows the calibration converged in eight iterations. The item parameter estimates are found in the table following ITEM PARAMETERS AFTER CYCLE 8. This table has the format seen in Chapter 4. The estimates for the five items are $\hat{\alpha}_1 = 1.226$, $\hat{\delta}_1 = -2.107$; $\hat{\alpha}_2 = 1.992$, $\hat{\delta}_2 = -0.499$; . . . ; $\hat{\alpha}_5 = 0.983$, and $\hat{\delta}_5 = 0.560$.¹³ (The final column of the item parameter estimates table contains the chi-square statistics that, as discussed in Chapter 4, should be ignored because of the length of our instrument.) In terms of item discrimination, all the estimated discrimination parameters are reasonably good items. As is the case with the 1PL model calibration, if one would like to measure individuals at the upper end of the θ continuum, it would be necessary to have items that are located beyond item 5. It can be seen that the mean discrimination of 1.459 ($SD = 0.381$) is slightly greater than the common $\hat{\alpha} = 1.421$ obtained with the 1PL model in Chapter 4. In effect, the 1PL model treats the item set as having a discriminating power approximately equal to the “mean discrimination” across items.

The top half of Table 5.3 contains the Phase 3 abridged output. Because BILOG is instructed to send the EAP $\hat{\theta}$ s to a separate file, the listing contains the line SCORES WRITTEN TO FILE MATH.SCO (i.e., the file name specified on the SAVE line). An example of the contents of this latter file is shown in the bottom half of Table 5.3.

The Phase 3 output is similar to that seen in Chapter 4. However, because of the NOPRINT subcommand on the SCORE line, the person table contains only information for the first three cases. The mean $\hat{\theta}$ for all 19,601 persons is -0.001 with an SD of 0.7882. Following these descriptive statistics, BILOG provides a summary of person parameter estimation accuracy. Although we can get a sense of the estimation accuracy at different points along the continuum (e.g., Figure 5.4), sometimes it is desirable to have a single bounded value that represents the quality of estimation for the entire continuum. One such index, the *empirical reliability*, is based on the ratio of the variance of the EAP $\hat{\theta}$ s to the sum of the variance of the $\hat{\theta}$ s and error variance (Zimowski et al., 2003). This index has a range from 0 to 1, with values that approach or are equal to 1.0 considered to be good values because they reflect small error variability. From Table 5.3 we have that the $\hat{\theta}$ VARIANCE is 0.6213 and the error variance of the EAP $\hat{\theta}$ s is 0.3852

TABLE 5.2. BILOG Output: Phases 1 and 2 (Abridged)

<Phase 1 results >							
:							
FILE ASSIGNMENT AND DISPOSITION							
=====							
SUBJECT DATA INPUT FILE MATH.DAT							
BILOG-MG MASTER DATA FILE MF.DAT							
CALIBRATION DATA FILE CF.DAT							
ITEM PARAMETERS FILE IF.DAT							
ITEM SCALE-SCORE FILE SF.DAT							
CASE WEIGHTING							
ITEM RESPONSE MODEL							
NONE EMPLOYED							
2 PARAMETER LOGISTIC							
LOGIT METRIC (I.E., D = 1.0)							
:							
<Phase 2 results begin>							
:							
DATA INPUT SPECIFICATIONS							
=====							
TYPE OF DATA							
MAXIMUM SAMPLE SIZE FOR ITEM CALIBRATION							
ALL SUBJECTS INCLUDED IN RUN							
SINGLE-SUBJECT DATA, NO CASE WEIGHTS							
20000							
:							
CYCLE 6; LARGEST CHANGE= 0.01593							
-2 LOG LIKELIHOOD = 110397.160							
CYCLE 7; LARGEST CHANGE= 0.00962							
[FULL NEWTON CYCLES]							
-2 LOG LIKELIHOOD: 110397.1034							
CYCLE 8; LARGEST CHANGE= 0.00305							
INTERVAL COUNTS FOR COMPUTATION OF ITEM CHI-SQUARES							

0. 691. 2857. 2600. 3500. 2714. 3854. 3385. 0.							

INTERVAL AVERAGE THETAS							

***** -1.964 -1.336 -0.767 -0.339 0.160 0.762 1.472*****							

SUBTEST MATH ; ITEM PARAMETERS AFTER CYCLE 8							
ITEM INTERCEPT SLOPE THRESHOLD LOADING ASYMPOTTE CHISQ DF							
S.E. S.E. S.E. S.E. S.E. (PROB)							

ITEM0001	2.584 0.044*	1.226 0.047*	-2.107 0.056*	0.775 0.030*	0.000 0.000*	164.8 (0.0000)	5.0
ITEM0002	0.995 0.032*	1.992 0.061*	-0.499 0.013*	0.894 0.027*	0.000 0.000*	2285.6 (0.0000)	3.0
ITEM0003	0.394 0.022*	1.551 0.041*	-0.254 0.014*	0.840 0.022*	0.000 0.000*	1174.8 (0.0000)	5.0
ITEM0004	-0.416 0.021*	1.544 0.039*	0.270 0.014*	0.839 0.021*	0.000 0.000*	362.8 (0.0000)	5.0
ITEM0005	-0.551 0.018*	0.983 0.026*	0.560 0.021*	0.701 0.019*	0.000 0.000*	1762.4 (0.0000)	5.0

* STANDARD ERROR							
LARGEST CHANGE = 0.003055							

5750.4 23.0 (0.0000)							

(continued)

TABLE 5.2. (*continued*)

NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE FOR TESTS WITH LESS THAN 20 ITEMS

PARAMETER	MEAN	STN DEV
SLOPE	1.459	0.381
LOG (SLOPE)	0.350	0.267
THRESHOLD	-0.406	1.039
:		

(i.e., VARIANCE of the ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS). Therefore, the

$$\text{EMPIRICAL RELIABILITY} = \frac{0.6213}{0.6213 + 0.3852} = 0.6173$$

This value may be considered marginally acceptable, given the instrument's length. This index is sometimes labeled "marginal reliability" in other calibration programs.¹⁴ Although a single accuracy value may be desirable, the trade-off is that it may potentially be misinterpreted. For instance, given the nonuniformity in the total information function (Figure 5.4), this empirical reliability understates the accuracy in the center of the metric and overstates the estimation accuracy for (approximately) $\hat{\theta}$ s < -1.5 and $\hat{\theta}$ s > 1.

The bottom half of Table 5.3 contains part of the MATH.SCO output file. As can be seen, the file contains only the title and the person estimate information; the first three lines match the cases displayed in the Phase 3 output. The layout of this file corresponds to the person table in the Phase 3 listing file. Recall that for the 2PL model the sufficient statistic for $\hat{\theta}$ is the weighted item responses, $\sum \alpha_j x_{ij}$. As a result, when item discrimination parameters (or their estimates) vary, different response patterns produce different $\hat{\theta}$ s. In other words, the pattern of responses is important because providing a response of 1 on an item with a larger α_j is more influential in estimating θ than if the item has a smaller α_j . For instance, the patterns 11110, 11101, 11011, 10111, and 01111 all have the same observed score (i.e., $X = 4$), but the corresponding $\hat{\theta}$ s are 0.7027, 0.4819, 0.4795, 0.3162, and 0.6047, respectively. In contrast, with the 1PL model all individuals with an $X = 4$ received the same $\hat{\theta}$ of 0.8238 (cf. Table 4.5) because the pattern of 1s and 0s is irrelevant. As we see, each estimate's SEE varies across the different response patterns for a given observed score (e.g., the $s_e(\hat{\theta})$); for 11110 it is 0.6404, but for 01111 it is 0.6284.

Fit Assessment: An Alternative Approach for Assessing Invariance

All of the previously discussed methods for examining model-data fit are appropriate for the 2PL model. For instance, in the previous two chapters we used fit statistics and empirical versus predicted IRF plots as part of our fit analysis. Moreover, we looked for

**TABLE 5.3. BILOG Phase 3 (Abridged) Output (Top Half)
and Abridged MATH.SCO Output File (Bottom Half)**

```

:
>SCORE MET=2, FIT,NOPRINT;
:
METHOD OF SCORING SUBJECTS: EXPECTATION A POSTERIORI
                               (EAP; BAYES ESTIMATES)
TYPE OF PRIOR: NORMAL
SCORES WRITTEN TO FILE MATH.SCO
SUBJECT FIT PROBABILITIES: YES
TYPE OF RESCALING: NONE REQUESTED
ITEM AND TEST INFORMATION: NONE REQUESTED
DOMAIN SCORE ESTIMATION: NONE REQUESTED

          QUAD
TEST     NAME  POINTS
-----
1      MATH    10
-----
:
GROUP   SUBJECT IDENTIFICATION
WEIGHT   TEST      TRIED   RIGHT   PERCENT      ABILITY      S.E.      MARGINAL
-----|-----|-----|-----|-----|-----|-----|-----|
1 00000
1.00 MATH      5       0       0.00 | -1.5821    0.6711 | 0.044414
1 10000
1.00 MATH      5       1      20.00 | -1.0752    0.6181 | 0.116346
1 01000
1.00 MATH      5       1      20.00 | -0.7957    0.5908 | 0.011625
-----|-----|-----|-----|-----|-----|-----|-----|
:
MEANS AND STANDARD DEVIATIONS OF SCORE ESTIMATES:

TEST:           MATH
MEAN:          -0.0010
S.D.:           0.7882
VARIANCE:       0.6213

ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

TEST:           MATH
RMS:            0.6206
VARIANCE:       0.3852

EMPIRICAL RELIABILITY: 0.6173
:

```

Abridged MATH.SCO output file

```

2PL item calibration

1 00000
1.00 MATH      5       0       0.00  -1.582075    0.671065  0.000000  0.044414
1 10000
1.00 MATH      5       1      20.00 -1.075171    0.618089  0.000000  0.116346
1 01000
1.00 MATH      5       1      20.00 -0.795662    0.590771  0.000000  0.011625
:
<MANY MORE PERSONS>
:
1 11000
1.00 MATH      5       2      40.00 -0.390327    0.566461  0.000000  0.074707
:
<MANY MORE PERSONS>
:
1 11100
1.00 MATH      5       3      60.00  0.125623    0.591210  0.000000  0.089643
:
```

(continued)

TABLE 5.3. (continued)

```

<MANY MORE PERSONS>
:
1 11110
1.00 MATH      5   4    80.00   0.702708   0.640394   0.000000   0.110770
1 11101
1.00 MATH      5   4    80.00   0.481892   0.615527   0.000000   0.069513
1 11011
1.00 MATH      5   4    80.00   0.479453   0.615300   0.000000   0.030829
1 10111
1.00 MATH      5   4    80.00   0.316185   0.602409   0.000000   0.014184
1 01111
1.00 MATH      5   4    80.00   0.604702   0.628386   0.000000   0.004111
:
<MANY MORE PERSONS>
:
1 11111
1.00 MATH      5   5   100.00   1.144288   0.702522   0.000000   0.157243

```

the evidence of estimate invariance by examining the correlation between the $\hat{\delta}$ s as well as the overlap among CIs from two random subsamples. With this latter approach, if our estimates are invariant within a linear transformation, then we have evidence supporting model-data fit. Previously, we created the random samples external to the calibration program. However, BILOG has the capacity to create random samples.

We begin by subdividing our calibration sample into two random subsamples. With BILOG we use its SAMPLE subcommand on the INPUT command line to specify that a subsample be taken from the calibration sample. For instance, to create the first subsample we might specify that 10,000 cases be randomly sampled from our calibration

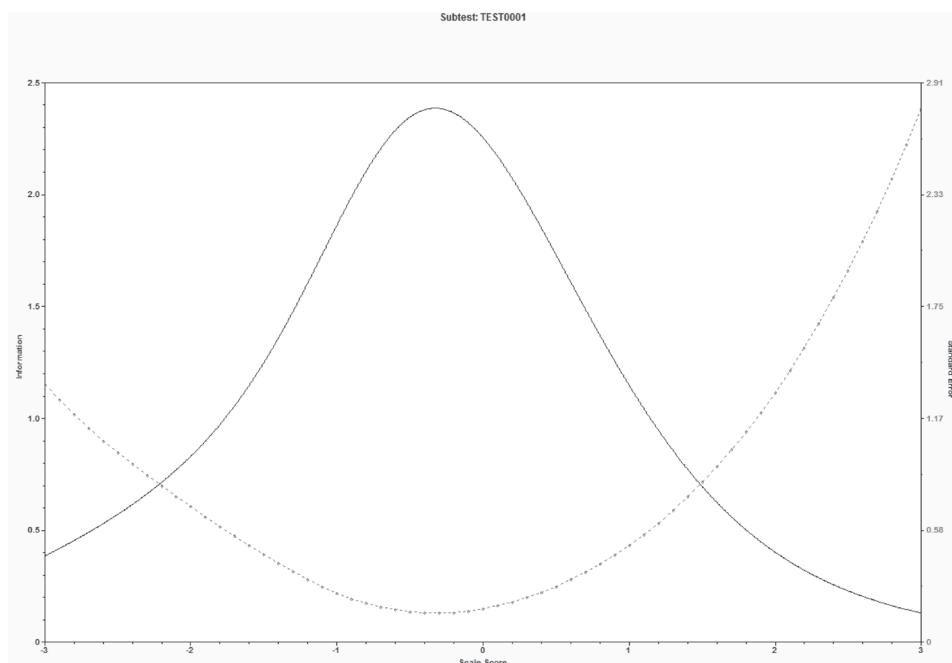


FIGURE 5.4. Total (test) information function

sample by adding the subcommand `SAMP = 10000` on the `INPUT` line. The second subsample is created by using the `ISEED` subcommand to specify the use of a different random number seed than the (default) seed value used for the first subsample. By using a different random number seed, a different random sample is generated. For example, the `INPUT` line `INPUT TOT = 5, NALT = 2, NIDC = 10, ISEED = 10, SAMP = 10000, TYPE = 1;` would generate a different random subsample of 10,000 individuals.¹⁵ Because BILOG performs the random sampling *in situ*, there is always the possibility that some cases will appear in multiple subsamples.

To implement this process, we create a command file for each subsample containing one of the above `INPUT` command lines. Furthermore, we save the item parameter estimates to separate files by using different file names on the `SAV` command line (`PARM = 'MATH1.PAR'` for subsample 1 and `PARM = 'MATH2.PAR'` for subsample 2). These files are used for our invariance analyses. After performing the separate calibrations, we examine the `DATA INPUT SPECIFICATIONS` section of each phase 1 output to verify that the correct number of cases is sampled (i.e., `MAXIMUM SAMPLE SIZE FOR ITEM CALIBRATION 10000`). After each subsample's calibration, we examine the item parameter estimate file.

With the two-parameter model, there are two sets of parameter estimates with which we need to be concerned, $\hat{\alpha}_j$ and $\hat{\delta}_j$. The correlation between the two samples' parameter estimates is 0.98532 for the $\hat{\alpha}_j$ s and 0.99998 for the $\hat{\delta}_j$ s. These correlations show that the item discrimination and location estimates from the two samples are highly linearly related. Therefore, given the magnitude of these correlations, we have evidence supporting model–data fit.

With the 2PL model, the correlation coefficients for the item discrimination and location estimates tell only part of the story because they do not fully reflect the interaction of these two types of parameters in describing an item. Therefore, we now present an additional (complementary) technique based on the difference in an item's response functions across subsamples. By using the IRF, we are able to see whether the item discrimination and location parameter interaction causes the two subsamples to vary.

We can assess the difference, or area, between two IRFs in multiple ways. Prior to their use, one should align or link the subsample metrics to one another by, for example, the total characteristic function equating approach (Chapter 11). One approach for examining the difference between two IRFs is to use the root mean squared difference (RMSD) between the probabilities represented by the two IRFs.¹⁶ In the current context, one set of probabilities (p_{js}) uses the parameter estimates from subsample s , and the other probability set (p_{jt}^*) uses the estimates from subsample t after transformation to the subsample s metric. The θ s used in calculating p_{js} and p_{jt}^* would be a series reflecting the desired degree of accuracy. For instance, one might use $-3.0, -2.95, \dots, 3.0$ (i.e., 0.05 logit difference) or whatever θ range is of particular interest.¹⁷ Decreasing the logit difference (e.g., from 0.05 to 0.005) will improve the index's accuracy as a measure of the difference between the two IRFs. The $RMSD_j$ statistic for item j is given by

$$RMSD_j = \sqrt{\frac{\sum (p_{js} - p_{jt}^*)^2}{n}}, \quad (5.9)$$

where n is the number of θ s used in calculating p_{js} and p_{jt}^* . As an example, if the range of θ is -3 to 3 in 0.05 increments, then $n = 121$. $RMSD_j$ has a range of 0 to 1 (inclusive), with small values indicating good agreement between the two IRFs; two identical IRFs have a $RMSD_j$ of 0.0. However, one should expect that even with perfect model–data fit, estimation error will be reflected in an item’s nonzero, albeit small, $RMSD_j$ value. From this perspective, a small $RMSD_j$ reflects two IRFs that may be considered to be sufficiently similar to one another and so not be a reason for concern.

In those cases where one observes a large $RMSD_j$ there may be various reasons for its magnitude. For instance, the item may be poorly written, or the model may have insufficient item parameters to accurately describe the item. Depending on the diagnosis of the cause(s) of the $RMSD_j$ magnitude, one may decide to omit the item from the instrument and retain only those items with small $RMSD_j$ values. $RMSD_j$ should be used in conjunction with a plot of the IRFs to determine whether the magnitude of the statistic is representative of a systematic difference across the continuum or reflects a difference for a particular portion of the continuum.

For our example, using the θ range -3 to 3 with a 0.05 increment, we obtain $RMSD_j$ s of 0.00759, 0.00144, 0.01180, 0.00163, and 0.00274 for items 1 through 5, respectively. For items 1, 2, 4, and 5, these $RMSD_j$ s represent small (trivial) differences between the two subsamples’ IRFs. Our largest $RMSD_j$ occurs for item 3 with a value of 0.01180. The corresponding IRFs plot (not presented) shows minor discrepancies below -0.5 and above 0.5; this item is slightly less discriminating and easier in subsample 2 than in subsample 1. Nevertheless, given that the correlations between the two subsamples’ estimates are high and that the five $RMSD_j$ s are small, we conclude that we have some evidence supporting the invariance of our estimates as well as model–data fit.

Raju (1990) offers a slightly more sophisticated approach for computing the area between two IRFs for the one-, two-, and three-parameter models as well as providing corresponding significance tests. However, given the sample sizes for some calibrations, the power associated with these statistical tests may render them not very meaningful; the null hypothesis is that there is no difference between the two IRFs. Moreover, if the subsamples contain cases in common, then the statistical test’s probability is not correct.

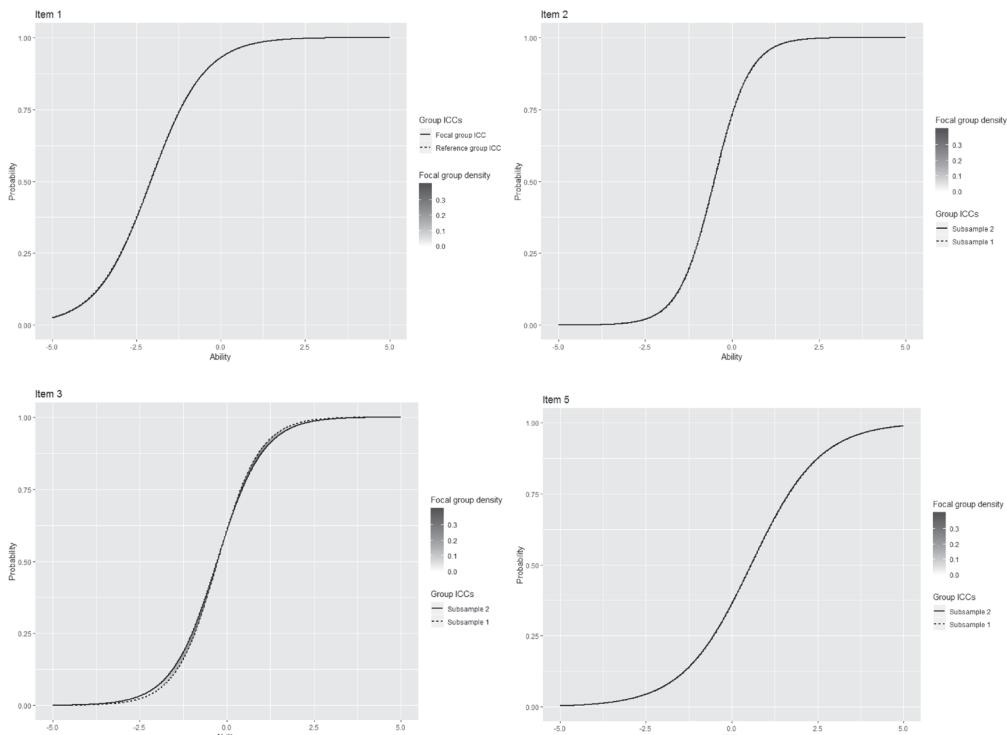
For the 2PL model the unsigned area, UA_{22} , between two IRFs for item j is obtained by

$$UA_{22j} = \left| \frac{2(\hat{\alpha}_{jt}^* - \hat{\alpha}_{js})}{\hat{\alpha}_{jt}^* \hat{\alpha}_{js}} \ln \left[1 + \exp \left(\frac{\hat{\alpha}_{jt}^* \hat{\alpha}_{js} (\hat{\delta}_{jt}^* - \hat{\delta}_{js})}{\hat{\alpha}_{jt}^* - \hat{\alpha}_{js}} \right) \right] - (\hat{\delta}_{jt}^* - \hat{\delta}_{js}) \right|, \quad (5.10)$$

where the asterisk reflects subsample t ’s estimates transformed to those of subsample s ; to simplify presentation, “ $\exp[z]$ ” is used instead of “ e^z .” Equation 5.10 simplifies to $|\hat{\delta}_{jt}^* - \hat{\delta}_{js}|$ when $\hat{\alpha}_{jt}^* = \hat{\alpha}_{js}$; Raju presents a version of UA_{22} for the three-parameter model (Chapter 6) for a common lower asymptote across subsamples.

For our instrument, and after transforming sample 2’s metric to that of sample 1, we obtain UA_{22j} values of 0.04802, 0.00758, 0.06898, 0.00772, and 0.01966 for items 1 through 5, respectively. Again, item 3 stands out as reflecting the greatest difference between the two samples’ IRFs. The agreement between UA_{22j} and $RMSD_j$ s is almost perfect ($r = 0.99574$).

To understand the magnitude of our UA_{22j} s, we can examine them graphically. By examining these graphs, we can diagnose the UA_{22j} values and determine whether we should be concerned. For example, Figure 5.5 shows the IRFs corresponding to the three largest and the smallest UA_{22j} s. As can be seen, items 1, 2, and 5 have IRFs that are virtually indistinguishable across the continuum. For item 3 with the largest UA_{223} , the discrepancy between the IRFs occur below approximately -1.25 and above approximately 1.25 , with the two IRFs crossing. However, the discrepancy is not very large. This item also demonstrates the unsigned nature of UA_{22j} . That is, a UA_{22j} value close to zero reflects IRFs that are identical and is not due to cancellation. For instance, we see that subsample 1's IRF is below that of subsample 2 (below approximately -1.25), but this difference is not canceled out by subsample 1's IRF being above subsample 2's (above approximately 1.25). Overall, the IRF discrepancies are small and within the margin of error. From the correlations, the $RMSD_j$ s, and the UA_{22j} we conclude that we have evidence of model–data fit. (The R packages DFIT [Cervantes, 2017a, 2017b] and difR [Magis, 2018; Magis, Béland, Tuerlinckx, & De Boeck, 2010] can be used to obtain Raju's area measures as well as other statistics. Endnote 18 shows how to do this with DFIT.)



^aFrom DFIT R session. An example function call:

```
> PlotNcdif(iItem = 5, itemParameters = twoPLitemests, irtModel = "2pl", focalIccText
= "Subsample 2", referenceIccText = "Subsample 1", main = "Item 5")
```

FIGURE 5.5. Comparison of subsample IRFs for items 1, 2, 3, and 5.

Example: Application of the 2PL Model to the Mathematics Data, MMLE, mirt

In Chapter 4 we show how to access and use `mirt` for parameter estimation. Consequently, below we assume that the data and the relevant libraries are loaded into our R workspace. To perform our calibration, we specify the 2PL as our `itemtype` in our call to the `mirt` function (`TwoPL = mirt(mathdata,1, '2PL', SE = T, SE.type = 'Fisher')`) and print the output object, `TwoPL`. Our calibration required 20 iterations to obtain convergence (Table 5.4).

Our $-2\ln L$ value is $-2(-55198.5) = 110,397$. From a data-fitting perspective, we can compare this value to that of the Rasch model. As shown in Table 4.6, we have $-2\ln L = 110,774.06$ for the Rasch model. The difference between two hierarchically related $-2\ln L$ s is known as a likelihood ratio and is evaluated with a chi-square distribution; the likelihood ratio is discussed further in Chapter 6. For our present purpose, we note that the difference of 377.06 is significant on 4 *dfs* and a critical $\chi^2_{\alpha=0.05,4} = 9.49$.¹⁹ As a result, our 2PL model fits significantly better than the Rasch model.

As mentioned above, `mirt` provides several information statistics. We can use our information criteria (e.g., AIC, BIC) in lieu of the above statistical test or to complement it. Our information criteria are concerned with model comparison and selection. In contrast to our above statistical test, these criteria can be used with nested or non-nested models.

In general, models with more parameters tend to fit a data set better than models with fewer parameters. Consequently, it is advisable to consider the additional parameters when examining model–data fit and to reduce the tendency toward model overparameterization. To this end, several information criteria are available. These statistics are typically based on an adjustment to the *deviance statistic*, $-2\ln L$. One commonly used measure is the *Akaike information criterion* (AIC) (Akaike, 1974) in which we adjust $-2\ln L$ for the number of parameters in the model

$$\text{AIC} = -2\ln L + 2\text{Nparm}, \quad (5.11)$$

where `Nparm` is the number of parameters being estimated. If the model has too many parameters relative to the sample size (i.e., `Nparm/N`), then AIC does not perform well (it tends to favor the model with the larger `Nparm`). To this end, Sugiura (1978) introduced an adjustment to AIC by adding a third term to Equation 5.11 to produce a *corrected AIC* (AICc). This term, $2\text{Nparm}(\text{Nparm} + 1)/(\text{N Nparm} - 1)$, is a bias correction factor. As `N` increases, the correction factor approaches zero and AICc essentially becomes AIC when we have large `N`. The model with the smallest AIC or AICc is the preferred model.

Schwarz (1978) introduced a different modification of AIC to create the *Bayesian information criterion* (BIC). The BIC statistic takes the number of parameters estimated by a model (`Nparm`), weights it by the transformed sample size (`N`), and thus “penalizes” the deviance statistic by taking into account the sample size. BIC is

$$\text{BIC} = -2\ln L + \ln(N)\text{Nparm}, \quad (5.12)$$

TABLE 5.4. mirt Session for the 2PL Calibration of the Mathematics Data

```

> # read data & load mirt

> print((TwoPL = mirt(mathdata,1,'2PL',SE=T,SE.type='Fisher')))
  Iteration: 20, Log-Lik: -55198.496, Max-Change: 0.00007

  Calculating information matrix.....

  Call:
  mirt(data = mathdata, model = 1, itemtype = "2PL", SE = T, SE.type = "Fisher")

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 20 EM iterations.
  mirt version: 1.30
  M-step optimizer: BFGS
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Information matrix estimated with method: Fisher
  Condition number of information matrix = 10.11284
  Second-order test: model is a possible local maximum

  Log-likelihood = -55198.5
  Estimated parameters: 10
  AIC = 110417; AICc = 110417
  BIC = 110495.8; SABIC = 110464
  G2 (21) = 395.33, p = 0
  RMSEA = 0.03, CFI = NaN, TLI = NaN
> M2(rasch,CI=0.95)      # Maydeu-Olivares & Joe statistic
      M2 df          p    RMSEA   RMSEA_2.5 RMSEA_97.5      SRMSR       TLI
stats 57.02138  5 5.005907e-11 0.02303976 0.01685003 0.02961753 0.01336253 0.9910097
      CFI
stats 0.9955048

> # calculate and display Orlando & Thissen's fit index using itemfit, but
> # increase precision of displayed values
> print(itemfit(TwoPL, fit_stats="S_X2"),digits=5)
      item     S_X2 df.S_X2 RMSEA.S_X2 p.S_X2
    1   I1 80.50442      2    0.04475 0.00000
    2   I2 19.40097      2    0.02107 0.00006
    3   I3 13.66416      2    0.01725 0.00108
    4   I4 65.60787      2    0.04028 0.00000
    5   I5 34.67477      2    0.02887 0.00000

> itemfit(TwoPL,S_X2.tables=T,empirical.table=1)                      # only item 1 is presented
$`theta = -1.2605`
      Observed  Expected z.Residual
cat_0        875    521.0757  15.504581
cat_1       1085   1438.9243 -9.330211

$`theta = -0.9915`
      Observed  Expected z.Residual
cat_0        635    403.2907  11.538103
cat_1       1325   1556.7093 -5.872726

$`theta = -0.5797`
      Observed  Expected z.Residual
cat_0        128    263.2686 -8.336763
cat_1       1832   1696.7314  3.283904

```

(continued)

TABLE 5.4. (*continued*)

```
$`theta = -0.3995`  
    Observed   Expected z.Residual  
cat_0      151  216.1997 -4.434226  
cat_1     1809 1743.8003  1.561338  
  
$`theta = -0.1141`  
    Observed   Expected z.Residual  
cat_0      314  156.7178 12.563781  
cat_1     1646 1803.2822 -3.703801  
  
$`theta = 0.1249`  
    Observed   Expected z.Residual  
cat_0       0  118.8432 -10.901521  
cat_1     1960 1841.1568  2.769675  
  
$`theta = 0.3908`  
    Observed   Expected z.Residual  
cat_0      63  86.82922 -2.5572715  
cat_1     1897 1873.17078  0.5505808  
  
$`theta = 0.6584`  
    Observed   Expected z.Residual  
cat_0      40  63.02143 -2.8999347  
cat_1    1920 1896.97857  0.5285683  
  
$`theta = 1.0245`  
    Observed   Expected z.Residual  
cat_0       0  40.43319 -6.3587098  
cat_1    1960 1919.56681  0.9228612  
  
$`theta = 1.146`  
    Observed   Expected z.Residual  
cat_0       0  34.87431 -5.905447  
cat_1    1961 1926.12569  0.794627  
  
> itemfit(TwoPL,group.bins=10,empirical.plot=1,empirical.CI=0) # produces Figure 5.6 (left)  
> itemfit(TwoPL,group.bins=6,empirical.plot=1,empirical.CI=0) # produces Figure 5.6 (right)  
  
> # examination of invariance -----  
> set.seed(99999) # set seed for random number generation  
> caseU=runif(19601) # 19601 x 5 variables  
> sortmathdata=mathdata # 19601 x 6 variables  
  
> sortmathdata$unif=caseU # 19601 x 6 variables  
  
> sortmathdata=sortmathdata[order(sortmathdata$unif),] # 9800 x 6 variables  
  
> mathdata1=sortmathdata[1:9800,] # 9800 x 6 variables  
> mathdata2=sortmathdata[9801:19601,] # 9801 x 6 variables  
  
> mathdata1=within(mathdata1,rm(unif)) # drop unif var; 9801 x 5  
> mathdata2=within(mathdata2,rm(unif)) # drop unif var; 9801 x 5  
  
> # subsample 1 calibration  
> TwoPL1 = mirt(mathdata1,1,'2PL',SE=T,SE.type='Fisher')  
  Iteration: 18, Log-Lik: -27557.717, Max-Change: 0.00008  
  
  Calculating information matrix...
```

(continued)

TABLE 5.4. (continued)

```

> coef(TwoPL1,simplify=TRUE,IRTpars=TRUE)
  $items
    a      b g u
  I1 1.212 -2.121 0 1
  I2 1.986 -0.519 0 1
  I3 1.585 -0.275 0 1
  I4 1.593  0.254 0 1
  I5 0.990  0.548 0 1

  $means
  F1
  0

  $cov
    F1
  F1  1

> # subsample 2 calibration
> TwoPL2 = mirt(mathdata2,1,'2PL',SE=T,SE.type='Fisher')
  Iteration: 18, Log-Lik: -27637.228, Max-Change: 0.00009

  Calculating information matrix.

> coef(TwoPL2,simplify=TRUE,IRTpars=TRUE)
  $items
    a      b g u
  I1 1.278 -2.035 0 1
  I2 2.058 -0.466 0 1
  I3 1.564 -0.226 0 1
  I4 1.542  0.277 0 1
  I5 1.002  0.557 0 1

  $means
  F1
  0

  $cov
    F1
  F1  1

> # extract item parameter estimates
> nitems = 5
> # subsample 1
> s1_itest=matrix(data=0,nrow= nitems,ncol=4)          # create a null matrix for the item est

> # subsample 2
> s2_itest=matrix(data=0,nrow= nitems,ncol=4)

> # subsample 1 extract each of the five est and save; there are alternate ways
> s1_itest=coef(TwoPL1,simplify=TRUE,IRTpars=TRUE)$items[,c('a','b')]
> s1_itest=as.data.frame(s1_itest)

> # subsample 2 extract each of the five est and save;
> s2_itest=coef(TwoPL2,simplify=TRUE,IRTpars=TRUE)$items[,c('a','b')]
> s2_itest=as.data.frame(s2_itest)

> cor(s1_itest$a,s2_itest$a)                            # correlation betw discr ests
[1] 0.9905938

> cor(s1_itest$b,s2_itest$b)                            # correlation betw locations ests
[1] 0.9999731

```

(continued)

TABLE 5.4. (*continued*)

```

> # subsample 1 mean & SD
> mean(s1_itest$a)
[1] 1.473113

> sd(s1_itest$a)
[1] 0.3848589

> mean(s1_itest$b)
[1] -0.4225517

> sd(s1_itest$b)
[1] 1.038773

> # subsample 2 mean & SD
> mean(s2_itest$a)
[1] 1.488647

> sd(s2_itest$a)
[1] 0.3917229

> mean(s2_itest$b)
[1] -0.3783626

> sd(s2_itest$b)
[1] 1.009974

> library(DFIT)

> # setting up meaningful variable names for next analysis
> colnames(s1_itest)=c('subsmpl_1.a','subsmpl_1.b')
> colnames(s2_itest)=c('subsmpl_2.a','subsmpl_2.b')

> s1_itest
  subsmpl_1.a subsmpl_1.b
 1   1.2115873  -2.1210137
 2   1.9863176  -0.5192444
 3   1.5848022  -0.2748753
 4   1.5932536   0.2542604
 5   0.9896021   0.5481145

> s2_itest
  subsmpl_2.a subsmpl_2.b
 1   1.277855  -2.0347895
 2   2.057859  -0.4659965
 3   1.563599  -0.2256354
 4   1.541987   0.2771868
 5   1.001935   0.5574215

> # deal w/ different metrics using mean-sigma (Ch 11); subsample 1 is target metric
> # obtain metric coefficients
> zeta=sd(s1_itest$subsmpl_1.b)/sd(s2_itest$subsmpl_2.b)

> kappa=mean(s1_itest$subsmpl_1.b)-zeta*mean(s2_itest$subsmpl_2.b)

> zeta
[1] 1.028514

> kappa
[1] -0.03340031

> # link metric
> s2_iteststar=s2_itest

```

(continued)

TABLE 5.4. (continued)

```

> s2_iteststar$subsmpl_2.a=s2_itest$subsmpl_2.a/zeta
> s2_iteststar$subsmpl_2.b=zeta*s2_itest$subsmpl_2.b + kappa

> s2_iteststar
  subsmpl_2.a subsmpl_2.b
I1    1.2424277 -2.1262106
I2    2.0008070 -0.5126844
I3    1.5202496 -0.2654696
I4    1.4992369  0.2516904
I5    0.9741572  0.5399158

> itests=cbind(s1_itest,s2_iteststar)                                # merge item estimate files
> itests
  subsmpl_1.a subsmpl_1.b subsmpl_2.a subsmpl_2.b
I1    1.2115873 -2.1210137  1.2424277 -2.1262106
I2    1.9863176 -0.5192444  2.0008070 -0.5126844
I3    1.5848022 -0.2748753  1.5202496 -0.2654696
I4    1.5932536  0.2542604  1.4992369  0.2516904
I5    0.9896021  0.5481145  0.9741572  0.5399158

> twoPLitests = list(focal = as.matrix(itests[, grep("subsmpl_1", names(itests))]),
                      reference = as.matrix(itests[, grep("subsmpl_2", names(itests))]))
> twoPLitests
  $focal
  subsmpl_1.a subsmpl_1.b
I1    1.2115873 -2.1210137
I2    1.9863176 -0.5192444
I3    1.5848022 -0.2748753
I4    1.5932536  0.2542604
I5    0.9896021  0.5481145

$reference
  subsmpl_2.a subsmpl_2.b
I1    1.2424277 -2.1262106
I2    2.0008070 -0.5126844
I3    1.5202496 -0.2654696
I4    1.4992369  0.2516904
I5    0.9741572  0.5399158

> # Raju unsigned area measure,
> UA_2PL = UnsignedArea(itemParameters = twoPLitests, irtModel = "2pl")
> UA_2PL
[1] 0.028730777 0.007676178 0.037964441 0.054605838 0.023247898

> # end of invariance check results --- 

> coef(TwoPL, IRTpars=TRUE, printSE=T)                                    # Item parameter estimates (N=19,601)
  $I1
    a      b      g      u
  par 1.245 -2.076  0   1
  SE  0.043  0.064 NA NA

  $I2
    a      b      g      u
  par 2.022 -0.492  0   1
  SE  0.047  0.013 NA NA

  $I3
    a      b      g      u
  par 1.574 -0.250  0   1
  SE  0.034  0.013 NA NA

```

(continued)

TABLE 5.4. (*continued*)

```
$I4
      a      b      g      u
par 1.567 0.266 0  1
SE  0.036 0.013 NA NA

$I5
      a      b      g      u
par 0.996 0.553 0  1
SE  0.024 0.019 NA NA

$GroupPars
      MEAN_1 COV_11
par      0      1
SE      NA     NA

> plot(TwoPL,type='trace' , theta_lim=c(-4,4))          # produces Figure 5.7 (top)
> plot(TwoPL, type = 'infotrace', theta_lim=c(-4,4))    # produces Figure 5.7 (bottom)

> # obtain EAP person estimates (fscores(TwoPL,method="EAP" ...)) & display
   first 6 cases
> head((peopleTwoPL=fscores(TwoPL,method="EAP",full.scores=T,full.scores.SE=T)),6)
      F1      SE_F1
[1,] -0.3861027 0.5666279
[2,]  0.1255866 0.5808742
[3,] -1.0728140 0.6101422
[4,]  0.1255866 0.5808742
[5,] -0.0256052 0.5723856
[6,]  0.1255866 0.5808742

> tail(peopleTwoPL,4)                                     # display last 4 cases
      F1      SE_F1
[19598,] 0.7015349 0.6387969
[19599,] 1.1460064 0.6997344
[19600,] 0.4760042 0.6121132
[19601,] 0.7015349 0.6387969

> mean(peopleTwoPL[,1])                                 # all 19,601 cases
[1] 1.365805e-06

> sd(peopleTwoPL[,1])                                  # all 19,601 cases
[1] 0.7874631

> # empirical reliability
> fscores(TwoPL,method="EAP",full.scores=T,full.scores.SE=T,returnER=T)
      F1
0.6200703

> # obtaining person fit info via personfit & display first 6 cases
> head((peopleTwoPLFit=personfit(TwoPL,method="EAP")), 6)
      Zh
1  1.0461586
2  1.0089647
3  1.0549464
4  1.0089647
5 -0.4792613
6  1.0089647

> tail(peopleTwoPLFit,4)
      Zh
19598  0.6534995
19599  0.7377341
19600 -0.3855811
19601  0.6534995
```

where N is the number of persons. The model with the smallest BIC indicates the model with the “best” comparative fit. As can be seen from Equation 5.12, the penalty increases as N and N_{parm} increase. Thus, BIC favors models with few parameters (“constrained” models). The *sample adjusted BIC* (SABIC; Schlove, 1987) replaces N in Equation 5.12 with $(N + 2)/24$ to reduce BIC’s sample size penalty. mirt will also calculate the *deviance information criterion* (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002); DIC is a modification of AIC and uses a data-based correction factor in lieu of N_{parm} as well as the posterior mean.

Our AIC and BIC are 110,417 and 110495.8, respectively. Again, from a data-fitting perspective, we can compare these to those of the Rasch model (Table 4.6). Both indices indicate the 2PL model fits the data relatively better than the Rasch model. This is, in fact, true using any of our information criteria; not surprisingly with a sample size of 19,601 $AIC_C = AIC$.

Although our significance test and information criteria indicate that the 2PL model is *preferred* to the Rasch model, this does not mean the 2PL model fits the data. To determine this we need to examine our other model-level and item-level fit information. As is the case with the Rasch calibration, our M_2 (57.021) is significant, but our RMSEA (0.02304) and its 95% CI, SRMR (0.01336), TLI (0.99101), and CFI (0.99550) all indicate close/good fit. As above, we take into account our sample size and do not give M_2 as much weight as our other indices. Thus, we interpret our RMSEA 95% CI, SRMR, TLI, and CFI as providing evidence of model–data fit.

We begin to examine item-level fit by using the `itemfit` function (`itemfit(TwoPL, stats = 'S_X2')`) to produce the fit statistic $S - X^2$ (Orlando & Thissen, 2000, 2003). $S - X^2$ uses the squared discrepancy between the observed and expected frequencies for each observed (summed) score group. That is, a manifest variable is used in lieu of categorizing a continuous IRT model-based latent variable estimate. These squared discrepancies are summed across score groups. This observed score-based index’s performance has, generally speaking, shown Type I error rates in line with the significance level, albeit with a concomitant increase in Type II error rates (Chon, Lee, & Dunbar, 2010; Orlando & Thissen, 2000, 2003; Stone & Zhang, 2003).²⁰

As can be seen, each of our items has a significant $S - X^2$; we opt to use $\alpha = 0.01$ to compensate for the large sample size’s effect on the test’s power. However, our items’ RMSEA indicates fit. To understand what is leading the $S - X^2$ s to be significant we inspect the residuals from the $S - X^2$ tables for each item using the `itemfit` function (e.g., for item 1: `itemfit(TwoPL, S_X2.tables = T, empirical.table = 1)`). We first look at each fractile for small expected frequencies (e.g., less than 4 or 5) because these could potentially inflate the statistic. Not surprisingly given our sample size, we do not find any. Then we focus on large `z.Residuals`. A negative sign tells us we are seeing fewer responses than expected, whereas a positive sign indicates more observed responses than expected. Comparing the observed and expected frequencies shows that we have fewer correct responses than expected below $\hat{\theta} = -0.9915$ and then (generally speaking) more correct responses than expected above $\hat{\theta} = -0.9915$, with a stronger agreement between what we observed and expect for a given fractile above $\hat{\theta} = 0.3908$. We can obtain a visual representation of this pattern by using the `empirical.plot` argument (i.e., `itemfit(. . . , empirical.plot = 1,`

. . .). As the left graph in Figure 5.6 shows, the predicted IRF tracks (more or less) the pattern of observed proportions.

Although categorizing the continuous latent math ability scale is necessary for calculating the itemfit chi-square statistics, the process discards information. As a result, these statistics should not be taken as the sole definitive statement of fit. Thus, we explore whether a reasonably close realistic agreement is possible by changing the number of fractiles in our plots from the 10 used with $S - X^2$ to 6 (i.e., the 6 possible X s). The right graph shows a closer agreement between the predicted and observed IRFs using `group.bins = 6`. We follow this process for the remaining items.

We proceed to obtain evidence of item parameter invariance using the procedure shown in Chapter 4 to generate random samples. After calibrating each sample, we obtain the correlations between the two sets of $\hat{\alpha}$ s (`cor(s1_itest$a, s2_itest$a)`) and between the two sets of $\hat{\delta}$ s (`cor(s1_itest$b, s2_itest$b)`). As can be seen, correlations with values above 0.99 provide evidence of invariance; the corresponding scatterplots reflect these correlations. Using DFIT, we obtain UA_{22} s for items 1 through 5 of 0.02873, 0.00768, 0.03796, 0.05461, and 0.02325, respectively. The corresponding plots of the samples' corresponding IRFs show that, overall, the IRF discrepancies are small and within the margin of error; these plots are similar to those in Figure 5.5. (These UA_{22} s differ from those found above with BILOG because the subsamples are not identical to those used with BILOG and the linking methods differ.) The foregoing model-data fit steps lead us to conclude that our estimates are a reasonable representation of our item-level data.

Our item parameter estimates (`coef(TwoPL, IRTpars = TRUE, printSE = T)`) are $\hat{\alpha}_1 = 1.245$, $\hat{\delta}_1 = -2.076$; $\hat{\alpha}_2 = 2.022$, $\hat{\delta}_2 = -0.492$; $\hat{\alpha}_3 = 1.574$, $\hat{\delta}_3 = -0.250$; $\hat{\alpha}_4 = 1.567$, $\hat{\delta}_4 = 0.266$; and $\hat{\alpha}_5 = 0.996$, $\hat{\delta}_5 = 0.553$. These estimates parallel those seen above with BILOG with $r_{\hat{\alpha}} = 1.0$ and $r_{\hat{\delta}} = 1.0$. Figure 5.7 shows the IRFs and item information functions for all five items. As was the case with the Rasch/1PL, the item information functions have their maxima at the item locations. As such, these maxima occur as different points on the continuum. However, in contrast to the Rasch/1PL, we see that the amount

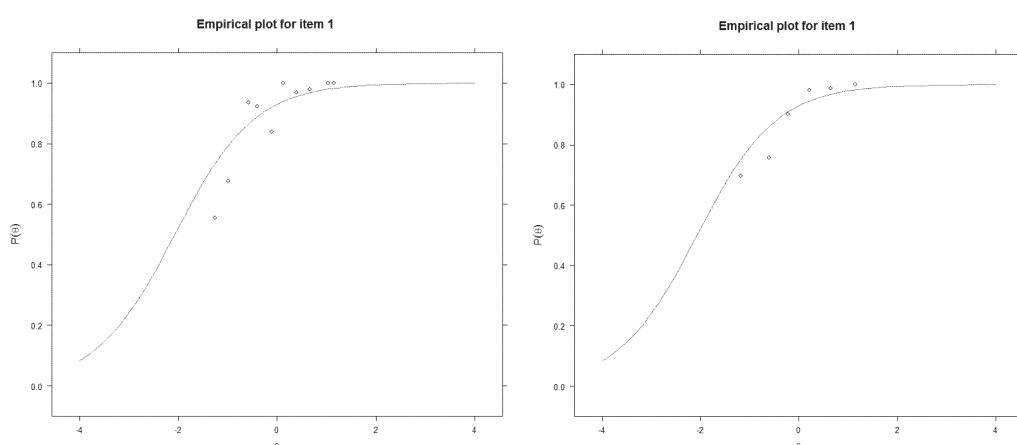


FIGURE 5.6. IRF for item 1 with observed proportions (left: 10 fractiles; right: 6 fractiles).

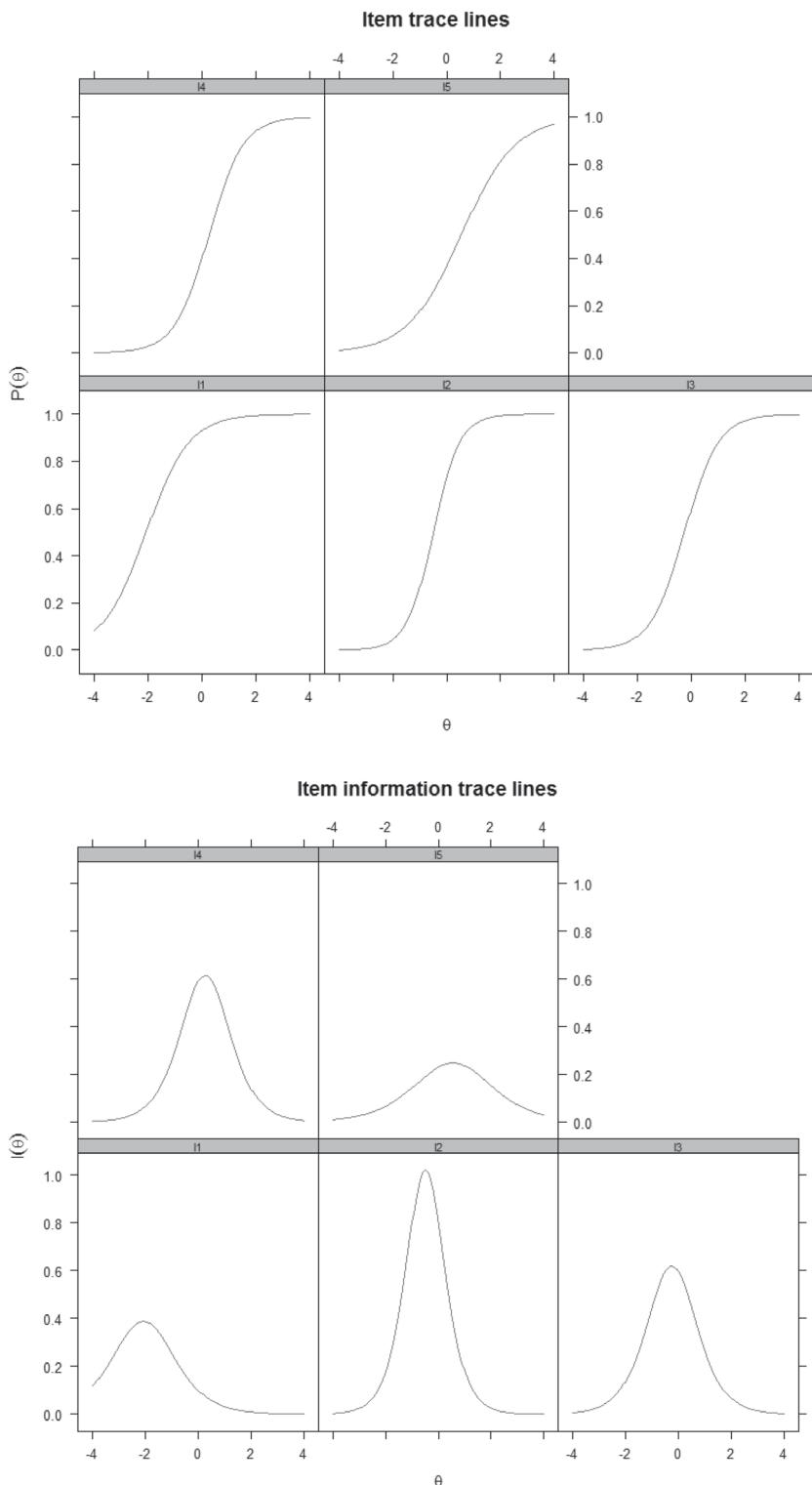


FIGURE 5.7. IRFs and item information for all items.

of information for estimating θ varies from item to item, with item 2 ($\hat{\alpha}_2 = 2.022$) providing the most information and item 5 ($\hat{\alpha}_5 = 0.996$) providing the least.²¹

As above, we use the `fscores` function to obtain our EAP $\hat{\theta}$ s (F1) with their corresponding standard errors ($PSD(\hat{\theta}) = SE_{\hat{\theta}}$ _ F1). For example, our first individual is estimated to be located at $\hat{\theta}_1 = -0.69603477$ with a standard error of estimate of 0.8589810. The empirical reliability is the average estimated conditional reliability. That is, for each individual we calculate an estimated conditional reliability $\hat{\rho}_{\theta} = 1 - PSD(\hat{\theta})$ whose average $\bar{\rho}_{\theta}$ is 0.620. This value may be considered marginally acceptable, given the instrument's length; our above interpretation of its usefulness applies here. To examine person fit, we use the `personfit` function with the fit statistic labeled " Z_h " (Drasgow, Levine, & Williams, 1985). However, because we have dichotomous data, the statistic is not " Z_h " but Z_3 (Drasgow et al., 1985); Z_h is for polytomous data (i.e., more than two responses). In general, values larger than 0 indicate a better fit than expected, given the 2PL model and values less than 0 indicating worse fit than expected. Although we defer further discussion of person fit until Chapter 6, suffice it to say that very large positive or very large negative " Z_h " values identify individuals whose response vectors should be examined. Moreover, because Z_h is not distributed $N(0,1)$ when $\hat{\theta}$ is used in its calculation, its false positive rates are inconsistent with the standard unit normal distribution based screening values (e.g., ± 1.96); factors that adversely affect the accuracy of parameter estimation (e.g., test length, model data fit, amount of missing data) can adversely affect Z_h . The Z_h s for the cases shown do not indicate any reason for person fit concern.

Information and Relative Efficiency

Above we mentioned that the mean α for the 2PL model calibration is greater than that of the 1PL model's α . The larger mean α for the 2PL model is reflected in an overall increase in the information available for estimating person locations. This increase may be represented either graphically (e.g., Figures 5.2 and 5.4) or in terms of the *total information area* index, I_A . We introduce the I_A as a complementary way to summarize in a single number the total information available for estimation. The total information area index represents the area under the total information function. Because the items contribute independently to the total information function, the area under the total information function is the sum of the item information areas

$$I_A = \sum_j^L \alpha_j I_j = \sum_j^L I_{A_j} \quad (5.13)$$

and the average total information area, \bar{I}_A , is

$$\bar{I}_A = \frac{I_A}{L}, \quad (5.14)$$

where

$$\iota = \frac{\chi_j \ln(\chi_j) + 1 - \chi_j}{1 - \chi_j}$$

and χ_j is the IRF's lower asymptote. Furthermore, whenever $\chi_j = 0$, then $\iota \equiv 1$. Because for both the 1PL and 2PL models $\chi_j = 0$ and $\iota = 1$, the *item information area* index (i.e., the area under the item information function) is

$$I_{A_j} = \alpha_j \iota. \quad (5.15)$$

Equations 5.13 to 5.15 provide a set of numerical indices that succinctly and non-graphically summarize the amount of information available for estimation without one having to estimate person locations or make distributional assumptions.²² All other things being equal, larger values of these indices indicate more information than do smaller values.

For instance, after linking the 1PL model's metric to that of the 2PL model, the 1PL model $\hat{\alpha} = 1.40$ and $I_A = L^* \hat{\alpha} = 5(1.40) = 7.0$. In contrast, for the 2PL model the area under the total information function is $I_A = \sum \hat{\alpha}_j = 1.226 + 1.992 + 1.551 + 1.544 + 0.983 = 7.296$. Therefore, for these data and metric, the 2PL model calibration of this instrument results in about 4% more total information for person estimation than does the 1PL model calibration (i.e., $7.296/7.0 = 1.0423$).

We may examine the total information functions for the 1PL and 2PL models to see how the distributions of the total information compare with one another. For convenience the 1PL model's total information function is superimposed over that of the 2PL model in Figure 5.8. First, we see that the maximum of the total information with the 1PL model is approximately 1.99 at $\theta \approx -0.15$. However, with the 2PL model the maximum of the total information is about 2.456 at $\theta \approx -0.30$. Second, as can be seen, the 2PL model is able to provide more information than the 1PL model in the approximate range of -1.4 to 0.55.²³

An alternative way of simultaneously comparing multiple information functions is by using a *relative efficiency* plot. Lord (1980; also see Birnbaum, 1968) presents the relative efficiency (RE) of one score x to another score y as the ratio of their information functions

$$RE\{x, y\} = \frac{I(\theta, x)}{I(\theta, y)}. \quad (5.16)$$

The scores x and y may come from two different instruments that measure the same construct θ , or they may arise from scoring the same instrument in two different ways. In the current context, the same instrument is scored in two different ways (i.e., the 1PL and 2PL models). Figure 5.9 contains the RE plot for the 1PL and 2PL models' calibration of the mathematics instrument. It is created by taking the 2PL model's total information and dividing it by the 1PL model's total information at each θ and plotting this ratio as a function of θ ; that is, $x = 2PL$, $y = 1PL$, and $RE\{2PL, 1PL\} = I(\theta, 2PL)/I(\theta, 1PL)$.

Figure 5.9 shows that the 2PL model is able to provide more information than the 1PL model in the approximate range of -1.4 to 0.55. In addition, at about -0.45 the 2PL model provides about 20% more information than the 1PL model, but outside the range

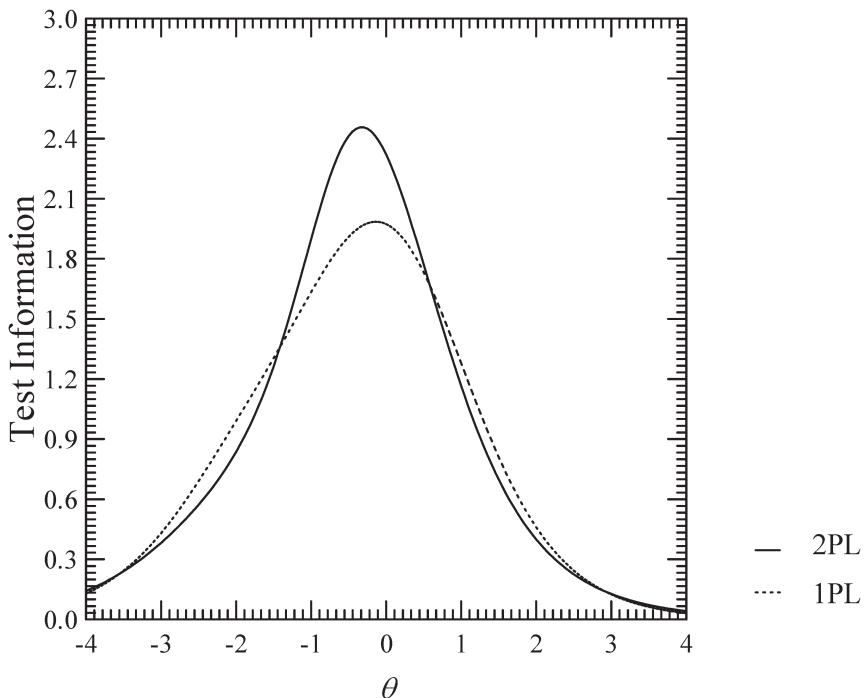


FIGURE 5.8. 1PL and 2PL models' total information functions.

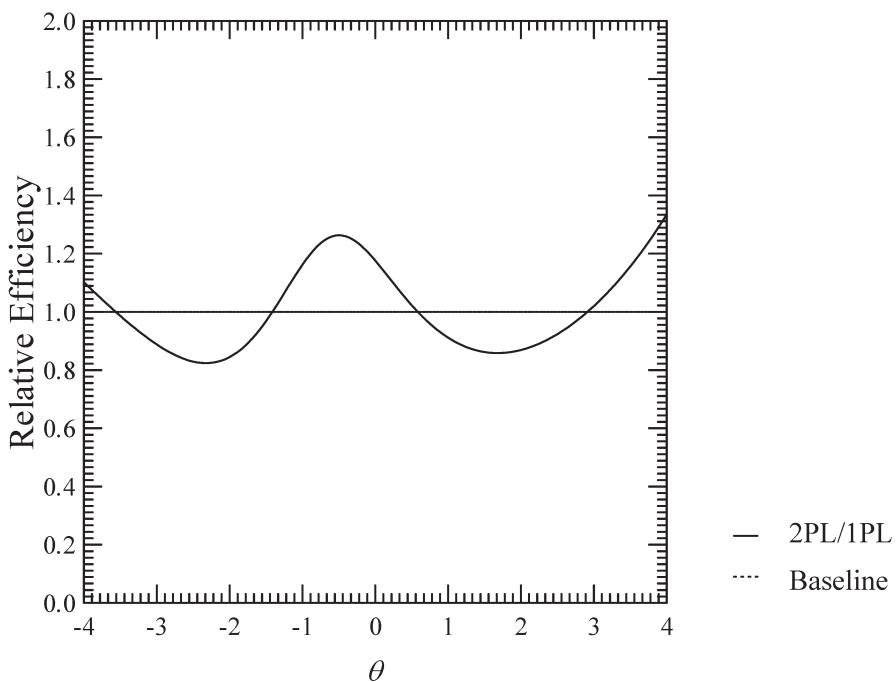


FIGURE 5.9. Relative efficiency plot for 1PL and 2PL models' calibrations.

of -1.4 to 0.55 the 1PL model provides more information than the 2PL model except at the extremes of the continuum. As the name would imply, the plot reflects the relative information as a function of θ . Therefore, one caveat in using *RE* plots is that, although the magnitude of information may be quite small at a given θ , this may not necessarily be evident from the plot. For instance, if the 2PL model provides 0.05 information for estimating a person located at $\theta = 1.5$ and the 1PL model provides 0.025 at this same point, then the 2PL model provides twice as much information as the 1PL model at this θ , although both provide a negligible amount of information at $\theta = 1.5$. This is what is happening in our case. Specifically, the information provided by both models differs at the extremes of the continuum: 0.04 vs. 0.03 at $\theta = 4$ and 0.14 vs. 0.13 at $\theta = -4$, with the large values in each pair associated with the 2PL.

Given that Figures 5.8 and 5.9 present similar information, one might ask, “What is the advantage of using a relative efficiency plot?” To help in understanding the benefit of an *RE* plot, the θ metric is transformed to be located at 25 by applying Equations 5.6 and 5.8 and letting $\kappa = 25$ and $\zeta = 5$.²⁴

The effect of this linear transformation on the metric is shown in Figure 5.10. This linear transformation does not affect the IRFs. Moreover, the shape of the information function is not affected by the linear transformation. However, it can be seen that the magnitude of information is affected by the transformation because the transformation affects the magnitude of the $\hat{\alpha}_j$ s. For example, for the 1PL model the maximum information is 1.985 (Figure 5.8) at $\theta = -0.15$, whereas on the transformed metric the maximum information becomes 0.079 at 24.85 . Therefore, given the relationship between information and the metric-dependent standard error of estimate, the information functions’ values are also metric dependent. However, the corresponding *RE* plot for the transformed metric (Figure 5.11) is identical to that shown in Figure 5.9 except that the abscissa reflects the transformed metric. Therefore, the benefit of an *RE* plot lies in its metric independence. In fact, relative efficiency is metric independent under any *monotonic* transformation of the metric (Lord, 1980). (See Appendix G, “Relative Efficiency, Monotonicity, and Information,” for more information on *RE* plots.) Moreover, the ratio of the total information area indices still shows that the 2PL model provides about 23% more total information than does the 1PL model; that is, $I(\theta, \text{2PL})_{\max}/I(\theta, \text{1PL})_{\max} = 0.09826/0.07941 = 1.2374$.

Summary

Items may vary in their capacity to discriminate among respondents located at different points along a variable’s continuum. Therefore, the use of an IRT model that captures this discrimination information may be useful for estimating a person’s location. The two-parameter model is one such model because it allows for items to vary not only in their locations, but also in their capacity to differentiate among persons located at different points on the continuum.

Because the discrimination parameter, α_j , is proportional to the slope of the IRF at the point of inflexion, the most obvious manifestation of items with different discrimi-

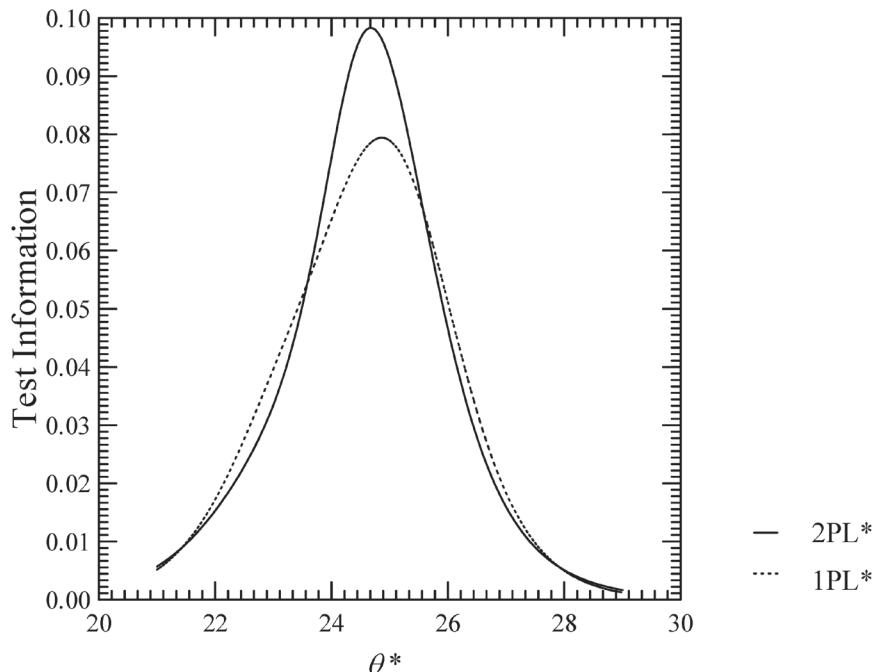


FIGURE 5.10. 1PL and 2PL models' total information functions for linearly transformed scale.

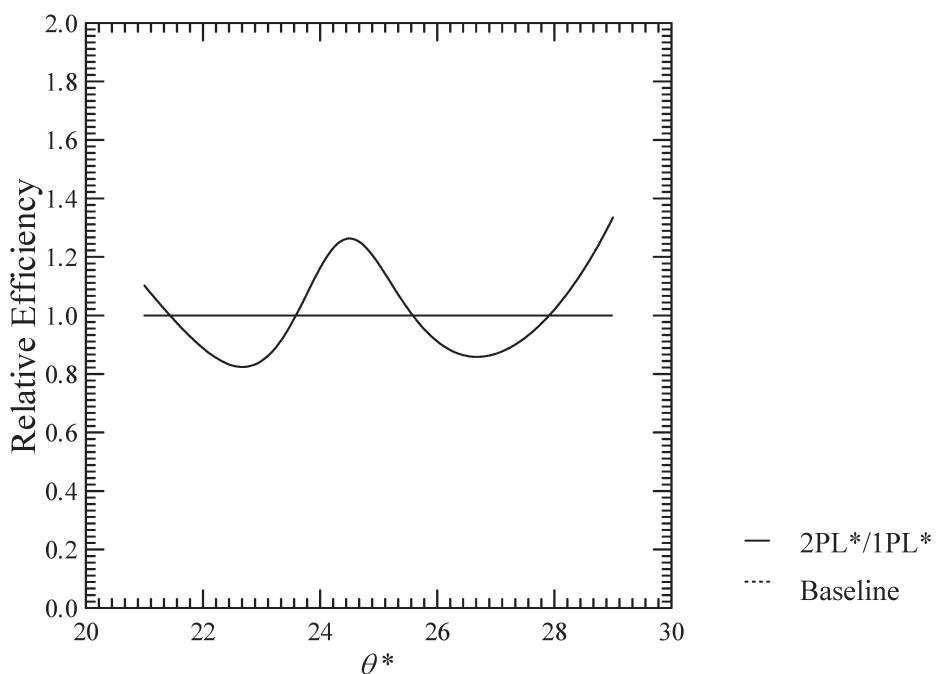


FIGURE 5.11. Relative efficiency plot for linearly transformed scale.

nation parameters is that their corresponding IRFs cross somewhere on the continuum. Although the discrimination parameter can theoretically vary from $-\infty$ to ∞ , items with positive α_j s between approximately 0.8 and 2.5 are considered good values. For the 2PL model, an IRF's point of inflection occurs at the item's location δ_j , and the slope at this point is $\alpha_j/0.25$.

As is the case with the 1PL model, the 2PL model has sufficient statistics. For estimating a person's location, the weighted sum of the item responses, $\sum \alpha_j x_{ij}$, is the sufficient statistic. This weighted sum provides more information than the unweighted sum used with the Rasch model except when the two sums are identical or proportional or when the $\hat{\alpha}$ s are not accurate. However, unlike the sufficient statistics for the 1PL model, the sufficient statistics for the 2PL model depend on unknown parameters.

A simple (albeit incomplete) method to assess the invariance of the parameter estimates at an instrument level is by computing correlations between the parameter estimates across two subsamples. A complementary strategy is to examine the difference between IRFs using the parameter estimates from the two subsamples. One index of this difference is $RMSD_j$; $RMSD_j$ should be used in conjunction with the corresponding plots of the IRFs.

With the two-parameter model, an item provides its maximum information at δ_j . In contrast to the one-parameter model, with the two-parameter model the maximum amount of item information may vary from item to item as α varies across items. To compare multiple information functions, one can use the relative efficiency plot. This plot can be created by taking the total information of one instrument and dividing it by the total information of the other instrument at each θ and plotting this ratio as a function of θ . In addition, the total information area index is introduced to complement graphical information function representations. This index summarizes in a single value the amount of information that an instrument provides and that is represented in its information function graph.

The 1PL and 2PL models share the constraint that their IRFs' lower asymptotes need to be zero. In Chapter 6 we relax this constraint and obtain the three-parameter model. With this model there is a parameter for the item's location, another for its discrimination capacity, and a third for the IRF's lower asymptote. This lower asymptote parameter is called the item's *pseudo-guessing parameter*, χ_j . Allowing the pseudo-guessing parameter to be nonzero permits the modeling of situations where one observes chance success on an item by persons of very low θ s. As such, the number of situations in which one may obtain model–data fit is potentially greater with the three-parameter model than with either the two- or one-parameter models.

Notes

1. Successful Rasch model–data fit indicates a one-to-one correspondence between mathematical operations, such as addition or subtraction of the measured values and the “structure of the properties of the objects that are measured” (Andrich, 1988, p. 17). (This correspondence is the gist of the concept of *fundamental measurement*.)

ment.) From this perspective the Rasch model is “chosen to express our intentions and not to simply describe the data” (Andrich, 1988, p. 14). Use of the model is predicated on the assumption that only items with empirical IRFs that have *approximately* the same slope are useful for estimating person locations. That is, although the ideal under the Rasch model is an item set with a constant item discrimination, in practice a certain degree of variability from this ideal is expected and tolerated. In point of fact, even though the biserials for the mathematics data vary from 0.397 to 0.564, we still have an acceptable degree of model–data fit for the Rasch model.

2. The 2PL model is, in effect, the ordinary logistic regression of the observed dichotomous responses on the latent person location and the latent item characterizations.
3. To determine the relationship between α_j and the IRF’s slope, we note that the slope is the same as the 2PL model’s first derivative, $p'_j = \alpha_j p_j(1 - p_j)$. By substitution of Equation 5.1 for p_j we have that the slope is

$$p'_j = \alpha_j \frac{e^{\alpha_j(\theta-\delta_j)}}{(1 + e^{\alpha_j(\theta-\delta_j)})(1 + e^{\alpha_j(\theta-\delta_j)})} = \alpha_j \frac{e^{\alpha_j(\theta-\delta_j)}}{(1 + e^{\alpha_j(\theta-\delta_j)})^2}. \quad (5.17)$$

Because α_j is defined at $\theta = \delta$, the slope is

$$p'_j = \alpha_j \frac{e^{\alpha_j(0)}}{(1 + e^{\alpha_j(0)})^2} = \frac{1}{4} \alpha_j = 0.25\alpha_j. \quad (5.18)$$

Therefore, α_j is proportional to the slope of the tangent line to the IRF at δ_j and may be thought of as a rate of change indicator for the IRF in the neighborhood around δ_j . When using the D scaling constant, the slope for the 2PL model is $0.25D\alpha_j = 0.425\alpha_j$.

As mentioned above, δ_j is defined as the IRF’s point of inflexion. Stated another way, this point is where the IRF switches from concave up to concave down. This is the point at which the rate of change of the slope (p'_j) changes from increasing to decreasing (or decreasing to increasing). To determine whether a function is increasing or decreasing, we examine the derivative’s sign. That is, to decide whether p'_j is increasing/decreasing, we examine its derivative’s sign. The derivative of p'_j ’s derivative (i.e., the derivative of p'_j) is p'_j ’s second derivative (i.e., p''_j). The second derivative for the 2PL model is $p''_j = \alpha_j^2 p_j(1 - p_j)(1 - 2p_j)$. As an example, assume an item is located at -0.25 and has an $\alpha = 1.574$.

θ	p_j	p'_j	p''_j	
-0.500	0.4029	0.3787	0.1158	concave up (convex)
-0.450	0.4219	0.3839	0.0943	"
-0.400	0.4412	0.3881	0.0718	"
-0.350	0.4607	0.3911	0.0483	"
-0.300	0.4803	0.3929	0.0243	"
-0.250	0.5000	0.3935	0.0000	inflexion point
-0.200	0.5197	0.3929	-0.0243	concave down

-0.150	0.5393	0.3911	-0.0483	"
-0.100	0.5588	0.3881	-0.0718	"
-0.050	0.5781	0.3839	-0.0943	"
0.000	0.5971	0.3787	-0.1158	"

We can see that as we approach the maximum slope (p'_j), the change in p'_j gets progressively smaller until we reach its maximum value. Subsequently, the change in p'_j gets progressively larger (albeit with a negative sign). Corresponding to these changes, we see that p''_j starts with positive values that get progressively smaller until it reaches a point of no change. Subsequently, p''_j turns negative and gets progressively larger. The point at which $p''_j = 0$ is when $\theta = \delta_j$ (i.e., where we have $p_j = 0.50$). By rearranging Equation 5.18, $\alpha = p'_j/0.25 = 0.3935/0.25 = 1.574$. In other words, $p''_j = 0$ when $p_j = (1 - p_j) = 0.50$. (Obviously, p''_j also equals 0 when $p_j = 1$ or $p_j = 0$. However, we ignore these degenerative solutions because for $p_j = 1$, then θ must equal ∞ or for $p_j = 0$, then θ must equal $-\infty$ [i.e., impossibilities].) Consequently, the point of inflexion ($p''_j = 0$) occurs at $p_j = 0.50$, and in IRT its location is symbolized δ_j .

4. As the distance between the person's and item's locations increases, a large α_j can lead to a loss of information away from δ_j because most of the information is concentrated in a neighborhood around δ_j . As α_j increases, the effective size of this neighborhood gets progressively smaller. For instance, in Figure 5.2 item 2 ($\alpha_2 = 1.5$) provides more information than does item 5 ($\alpha_5 = 3.0$) below approximately -0.05 . This "attenuation paradox" (Lord & Novick, 1968) can be summarized as follows: If each α_j is extremely high, then one has virtually error-free discrimination between θ levels in a small neighborhood around each δ_j , but as one leaves this neighborhood, there is "virtually no other information for discrimination or estimation" (p. 465).
 5. The area under $I_j(\theta)$ is equal to α_j (with the D scaling constant the area is equal to $D\alpha_j$; Lord & Novick, 1968). That is,
- $$\int_{-\infty}^{\infty} I_j(\theta) d\theta = \alpha_j.$$
- Moreover, the area under $\sqrt{I_j(\theta)}$ is equal to π (Samejima, 1994).
6. The idea of "how quickly the IRF is changing" is captured by the slope of the IRF and is reflected in the model's first derivative, p'_j , whereas $p_j(1 - p_j)$ is the variance at θ .
 7. The standard error for the person location estimate under the 2PL is

$$s_e(\theta) = \frac{1}{\sqrt{\sum_j^L \alpha_j^2 p_j(1 - p_j)}}, \quad (5.19)$$

where p_j is conditional on $\hat{\theta}_i$.

8. The observed log likelihood values can be obtained by executing the R program:

```
# assumes data reside in x
item = 1      # set item of interest
```

```
# general initializations
minAlpha=-1.0; maxAlpha=4.0; minDelta=-3.0; maxDelta=3.0; incr =
0.1
nalphaPts = (abs(minAlpha)+abs(maxAlpha))/incr+1 # number of unique
values
ndeltaPts = (abs(minDelta)+abs(maxDelta))/incr+1 # number of unique
values

N = length(x[,item]) # N persons
X = rowSums(x) # calc observed scores, X
z = scale(x,center=T,scale=T) # convert X to z

delta=seq(minDelta,maxDelta,incr) # create deltas

alpha=seq(minAlpha,maxAlpha,incr) # create alphas

lnL = rep(0.0,(nalphaPts*ndeltaPts))
g = 1
for (j in 1:ndeltaPts) {
  for (k in 1:nalphaPts) {
    lnLike = 0.0
    for (i in 1:N) {
      p = 1/(1+exp(-1.0*alpha[k]*(z[i]-delta[j])))
      lnLike = lnLike + x[i,item]*log(p) +
      (1-x[i,item])*log(1 - p)
    } # for i loop

    lnL[g] = lnLike
    g = g + 1

  } # for k loop
} # for j loop

# output- 3 vectors: alpha, delta, and lnL
# For example, write.csv((cbind((cbind(alpha,delta)),lnL)),"<filename.
csv")
```

Note: (1) Calculating the log likelihood values with a large data set (e.g., math-data's 19,601 cases) can be time consuming. Depending on the computer, it can take 18–45 minutes. (2) To use R's `persp` function requires that calculating the log likelihood (`lnL`) be modified so that `lnL` is a matrix.

9. The sufficient statistic for estimating a person's location is the weighted composite $\sum \alpha_j x_{ij}$, where the weights are given by the item discrimination parameters. From a more general perspective, one can write the weighted composite as $\sum w_j x_{ij}$. As a result, one might ask, "Is there a weight that will provide more information for estimating a person's location than that provided by α_j ?" Lord (1980) shows that the information function for the weighted composite is the total information function and that this is the maximum information attainable by any scoring method. Therefore, the optimal scoring weight for an item j is

$$w_j(\theta) = \frac{p'_j}{p_j(1-p_j)}. \quad (5.20)$$

Because the first derivative for the 2PL model is

$$p'_j = \alpha_j p_j (1 - p_j),$$

then by substitution and simplification the optimal weight for the 2PL model is

$$w_j(\theta) = \frac{\alpha_j p_j (1 - p_j)}{p_j(1 - p_j)} = \alpha_j. \quad (5.21)$$

Therefore, α_j is the optimal weight for maximizing information for locating persons. As would be expected, for the 1PL model $w_j(\theta) = \alpha$ and for the Rasch model $w_j(\theta) = 1$. With the scaling factor D we have that $w_j(\theta) = D\alpha_j$, $w_j(\theta) = D\alpha$, and $w_j(\theta) = D$ for the 2PL, 1PL, and Rasch models, respectively.

10. In general, one no longer sees JMLE used for parameter estimation for the two-parameter model. Nevertheless, for completeness, we describe some of the parameter recovery research in this area. Hulin, Lissak, and Drasgow (1982), in a study of JMLE (specifically, LOGIST), conducted a Monte Carlo study to investigate the effects of four sample sizes ($N = 200, 500, 1000$, or 2000) and three instrument lengths ($L = 15, 30$, or 60 items) on the accuracy of parameter estimation. They found that the average error (i.e., root mean squared) for instruments of at least 30 items was no greater than 0.05 for a sample size of $1,000$ and less than 0.07 with 500 cases. Lim and Drasgow (1990) found that samples of 250 tended to result in greater biased item parameter estimates and standard errors than did the larger sample size of 750 persons. In general, JMLE parameter and standard error estimates were not as accurate as those of MMLE, even when using a 25-item instrument and a sample size of $1,000$. It may be conjectured that these results are due to the ping-pong nature of estimating persons and item parameters. That is, less accurate $\hat{\theta}$ s at one stage (e.g., due to an instrument's length) affect the accuracy of the $\hat{\alpha}$ s and $\hat{\delta}$ s, which then in turn affect the accuracy of the subsequent $\hat{\theta}$ s, and so on. In short, the estimation errors are compounded and/or propagated through each stage of estimating the persons and items. On the basis of these results, it appears that instruments of 25 items or more with sample sizes of at least $1,000$ persons should be used when using JMLE.
11. In addition to the unidimensionality assumption, we should also determine the tenability of the conditional independence assumption. We investigate this assumption's tenability for the mathematics data in Chapter 6. Although the approach we use is applicable for the 1PL and 2PL models, Glas (1999) has developed an alternative procedure specifically for the two-parameter model.
12. If we had used MLE for estimating the person locations, then we would not be able to obtain MLE estimates for individuals with zero scores or perfect scores. In

Chapter 3 two approaches were mentioned that could be used for these cases: the half-item rule and the addition of a constant (e.g., 0.5) to a zero score and its subtraction from a perfect score. With the two-parameter model, one could use an additional rule. Individuals with zero scores are assigned a value equal to $\hat{\alpha}_{\min}/2$ (i.e., $X = \hat{\alpha}_{\min}/2$), and for persons with perfect scores the observed score, X , equals $\sum \hat{\alpha} - \hat{\alpha}_{\max}/2$, where $\hat{\alpha}_{\min}$ and $\hat{\alpha}_{\max}$ are the minimum and maximum $\hat{\alpha}_j$ s, respectively, for the administered instrument.

13. When all individuals respond to an item in the same way (e.g., all responses are 1s), then it is not possible to obtain estimates of the item's parameters using either MLE or MMLE. In the case of MMLE, this demonstrates that MMLE is not a Bayesian procedure. However, if one requests that BILOG-MG (version 3) save the item parameter estimates to an external file (i.e., the "PAR" file), the item(s) with the zero variance will have item parameter "estimates" in the file (albeit with an extreme location estimate), although the Phase 2 output will not have an entry for these zero variance item(s). If there is a need to obtain estimates when all persons correctly respond to an item, then a kludge (i.e., work-around) is to randomly select a case and change the response to be an omitted response (e.g., changing the observed correct response to blank). (A similar approach is used with an item to which all responses are incorrect.) This allows the item's parameter(s) to be estimated. By randomly selecting the case, one can treat the "omitted" response as missing completely at random. In the case of the 1PL model, the location estimate may be "comparatively" extreme. With the 2PL model, the estimates may be reasonable, depending on the sample size.
14. Green et al. (1984) define the marginal reliability as

$$\rho = \frac{\sigma_{\theta}^2 - \sigma_{em}^2}{\sigma_{\theta}^2} \quad (5.22)$$

where σ_e^2 is the variance of the observed person locations, σ_{em}^2 is the marginal measurement error

$$\sigma_{em}^2 = \frac{\int_{-\infty}^{\infty} \sigma_e^2(\theta)g(\theta)d\theta}{\int_{-\infty}^{\infty} g(\theta)d\theta}$$

and $g(\theta)$ is the person distribution that, when $g(\theta)$ is normally distributed, can be evaluated using a Gaussian quadrature approach.

As an alternative to an index for the entire metric, one may calculate the *conditional reliability* (Green et al., 1984) at a θ point. They define conditional reliability as

$$\rho_{\theta} = \frac{\sigma_{\theta}^2 - \sigma_e^2(\hat{\theta})}{\sigma_{\theta}^2}, \quad (5.23)$$

where σ_e^2 is the variance of the observed person locations and $\sigma_e^2(\hat{\theta})$ is the expected variance error of estimate. The term ρ_θ specifies “the reliability if everyone were measured with the same precision as those persons” (Green et al., 1984, p. 353) located at θ . In addition, the graphing of ρ_θ as a function of θ would allow an assessment of the estimation properties of the instrument at various levels of θ on a bounded and potentially more easily interpreted (ordinate) scale from 0 to 1 than that used with the total information function.

An additional reliability index introduced for the EAP location estimate rests on the assumption that the latent variable is normally distributed in the population with a mean of 0 and a variance of 1 (Bock & Mislevy, 1982). Thus, with this standardized situation, Equation 5.23 simplifies to $\rho_\theta = 1 - PSD(\theta)^2$.

If we standardize our $\hat{\theta}$ s and adjust their corresponding $s_e(\hat{\theta})$ s by $s_e^*(\hat{\theta}) = s_e(\hat{\theta}) / s_{\hat{\theta}}$ where $s_{\hat{\theta}}$ is the standard deviation of the $\hat{\theta}$ s, then our estimated reliability is

$$\hat{\rho} = 1 - \text{avg}(PSD^*(\hat{\theta}_i)^2). \quad (5.24)$$

With EAP estimates $s_e(\hat{\theta}) = PSD(\hat{\theta}_i)$ and $s_e^*(\hat{\theta}) = PSD^*(\hat{\theta}_i)$, $\text{avg}(PSD^*(\hat{\theta}_i)^2)$ is the average of the variance errors. Equation 5.24 will provide the same estimate as

$$\hat{\rho} = \frac{s_{\hat{\theta}}^2}{s_{\hat{\theta}}^2 + \text{avg}(s_e^2(\hat{\theta}))}. \quad (5.25)$$

15. BILOG command file & abridged phase 2 output for subsample 1: 2PL item calibration; random sample 1

```
>GLOBAL DFNAME='MATH.DAT', NPARM=2, NWGHT=0, LOG,SAVE;
>SAV PARM='MATH.PAR', SCO='MATH.SCO';
>LENGTH NITEMS=5;
>INPUT NTOT=5,NALT=2,NIDC=10, ISEED=20,SAMP=10000,TYPE=1;
>ITEMS;
>TEST TNAMES='MATH',
      INumber = (1(1)5);
      (10A1,T1,5(1X,1A1))
>CALIB CYCLES=20, NEWTON=20, CHI=(5,9), PLOT=1.0;
>SCORE MET=2,NOPRINT;

:
ITEM      INTERCEPT      SLOPE      THRESHOLD      LOADING      ASYMPTOTE      CHISQ      DF
      S.E.          S.E.          S.E.          S.E.          S.E.          (PROB)
-----
-- 
ITEM0001 |  2.606 |  1.251 | -2.083 |  0.781 |  0.000 |  141.1 |  4.0
           | 0.062* | 0.065* | 0.076* | 0.041* | 0.000* | (0.0000)
           |         |         |         |         |         |
ITEM0002 |  1.023 |  1.945 | -0.526 |  0.889 |  0.000 |  185.0 |  4.0
           | 0.044* | 0.083* | 0.019* | 0.038* | 0.000* | (0.0000)
           |         |         |         |         |         |
ITEM0003 |  0.433 |  1.524 | -0.284 |  0.836 |  0.000 |  207.3 |  4.0
           | 0.030* | 0.056* | 0.020* | 0.031* | 0.000* | (0.0000)
```

ITEM0004	-0.392	1.530	0.256	0.837	0.000	124.5	4.0
	0.029*	0.054*	0.020*	0.030*	0.000*	(0.0000)	
ITEM0005	-0.555	1.000	0.555	0.707	0.000	227.4	4.0
	0.025*	0.038*	0.029*	0.027*	0.000*	(0.0000)	

--

* STANDARD ERROR

:

BILOG command file and abridged phase 2 output for subsample 2:

2PL item calibration; random sample 2

```
>GLOBAL DFNAME='MATH.DAT', NPARM=2, NWGHT=0, LOG,SAVE;
>SAV PARM='MATH.PAR', SCO='MATH.SCO';
>LENGTH NITEMS=5;
>INPUT NTOT=5,NALT=2,NIDC=10, ISEED=10,SAMP=10000, TYPE=1,
>ITEMS;
>TEST TNAMES='MATH',
    INumber = (1(1)5);
(10A1,T1,5(1X,1A1))
>CALIB CYCLES=20, NEWTON=20, CHI=(5,9), PLOT=1.0;
>SCORE MET=2,NOPRINT;
```

Output:

:

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
ITEM0001	2.607	1.243	-2.096	0.779	0.000	157.3	4.0
	0.063*	0.067*	0.078*	0.042*	0.000*	(0.0000)	
ITEM0002	0.997	1.948	-0.512	0.890	0.000	190.7	4.0
	0.043*	0.081*	0.019*	0.037*	0.000*	(0.0000)	
ITEM0003	0.441	1.668	-0.264	0.858	0.000	168.1	4.0
	0.032*	0.062*	0.019*	0.032*	0.000*	(0.0000)	
ITEM0004	-0.417	1.544	0.270	0.839	0.000	135.6	4.0
	0.030*	0.054*	0.020*	0.029*	0.000*	(0.0000)	
ITEM0005	-0.568	1.000	0.568	0.707	0.000	226.8	4.0
	0.026*	0.037*	0.029*	0.026*	0.000*	(0.0000)	

--

* STANDARD ERROR

These subsample 2 estimates are transformed to be on the same metric as subsample 1's (see Chapter 11). The equating coefficients are $\zeta = 1.0121$ and $\kappa = -0.0088$.

16. This approach is similar to the UA and SOS differential item functioning measures presented in Shepard, Camilli, and Averill (1981) as well as Shepard, Camilli, and Williams (1984).
17. The θ range used for the calculating RMSD should be the range of interest. For

instance, in a particular application, we might focus on a range of θ around a cut-point, θ' . Therefore, we would be most concerned with the similarity between IRFs within this range around θ' and less concerned with discrepancies that occur farther away from θ' .

18. We use the total characteristic function approach to link subsample 2's metric to that of subsample 1's; Chapter 11 shows how to do this. The R session to obtain UA_{22} using DFIT using the transformed sample 2 estimates is

```
> library(DFIT)
> # sample 1 discr/location (V1 & V2), sample 2 linked discr/location (V3 & V4)
> # subsample 1 is the target metric
> temp=read.table("BILOGitemEst.dat",header=F)
> temp
      V1        V2        V3        V4
1 1.25091 -2.08302  1.229 -2.130
2 1.94531 -0.52588  1.925 -0.527
3 1.52372 -0.28450  1.648 -0.276
4 1.53033  0.25634  1.526  0.264
5 1.00039  0.55513  0.988  0.566

> names(temp) = c("subsmpl_1.a","subsmpl_1.d","subsmpl_2.a","subsmpl_2.d")
> temp
      subsmpl_1.a subsmpl_1.d subsmpl_2.a subsmpl_2.d
1     1.25091    -2.08302     1.229    -2.130
2     1.94531    -0.52588     1.925    -0.527
3     1.52372    -0.28450     1.648    -0.276
4     1.53033     0.25634     1.526     0.264
5     1.00039     0.55513     0.988     0.566

> twoPLtests = list(focal = as.matrix(temp[, grep("subsmpl_1",
names(temp))]), reference = as.matrix(temp[, grep("subsmpl_2",
names(temp))]))

> twoPLtests
$ focal
      subsmpl_1.a subsmpl_1.b
[1,]     1.25091   -2.08302
[2,]     1.94531   -0.52588
[3,]     1.52372   -0.28450
[4,]     1.53033    0.25634
[5,]     1.00039    0.55513

$ reference
      subsmpl_2.a subsmpl_2.b
[1,]     1.229    -2.130
[2,]     1.925    -0.527
[3,]     1.648    -0.276
[4,]     1.526     0.264
[5,]     0.988     0.566
> # Raju unsigned area measure,
```

```
> UA_2PL = UnsignedArea(itemParameters = twoPLitests, irtModel =
  "2pl")
> UA_2PL
[1] 0.048015956 0.007576472 0.068975598 0.007719093 0.019664130

> # Raju signed area measure
> SA_2PL = SignedArea(itemParameters = twoPLitests, irtModel =
  "2pl")
> SA_2PL
     I1      I2      I3      I4      I5
-0.04698 -0.00112  0.00850  0.00766  0.01087
```

19. We can use the `anova` function with our output objects to perform this calculation.

For example, for our comparison of the Rasch and 2PL models we have

```
> anova(TwoPL,rasch)

Model 1: mirt(data = mathdata, model = 1, itemtype = "Rasch", SE
= T,
  SE.type = "Fisher")
Model 2: mirt(data = mathdata, model = 1, itemtype = "2PL", SE =
T,
  SE.type = "Fisher")

  AIC    AICc   SABIC      HQ      BIC    logLik      X2  df   p
1 110786.1 110786.1 110814.3 110801.5 110833.4 -55387.03      NaN NaN NaN
2 110417.0 110417.0 110464.0 110442.8 110495.8 -55198.50 377.066   4   0
```

The difference between the two $-2\ln L_s$ is found in the column labeled `X2` for chi-square. For a comparison of 1PL and 2PL models we have

```
> anova(TwoPL,OnePL)

Model 1: mirt(data = mathdata, model = OnePLconstr, SE = T, SE.type =
"Fisher")
Model 2: mirt(data = mathdata, model = 1, itemtype = "2PL", SE = T,
  SE.type = "Fisher")

  AIC    AICc   SABIC      HQ      BIC    logLik      X2  df   p
1 110786.1 110786.1 110814.4 110801.6 110833.4 -55387.06      NaN NaN NaN
2 110417.0 110417.0 110464.0 110442.8 110495.8 -55198.50 377.131   4   0
```

20. Stone (2000) introduced a chi-square statistic that uses resampling. This statistic, χ^{2*} , has shown good statistical performance. Specifically, with respect to power, χ^{2*} has performed better than $S - X^2$ while maintaining a nominal Type I error rate in line with the significance level (Chon, Lee, & Dunbar, 2010; Stone & Zhang, 2003). However, χ^{2*} can be computationally intensive. For this example, obtaining the following χ^{2*} results took several minutes.

```
> print(itemfit(TwoPL,fit.stats="X2*"),digits=5)
  item  X2_star p.X2_star
1   I1 24.02505      0
2   I2  8.83218      0
3   I3 14.37885      0
4   I4 20.34917      0
5   I5 18.38100      0
```

As can be seen, all five items show significant χ^{2*} s. Therefore, both $S - X^2$ and χ^{2*} indicate a lack of item-level misfit. χ^{2*} uses resampling in determining the probability. As a result, the probability for a χ^{2*} will most likely vary across repeated calls to calculate χ^{2*} . In this case, because the p -value is essentially zero to five places, repeated calls to calculate χ^{2*} will not show the variability in p -values. In Chapter 7 we use the rescaled version of χ^{2*} , χ_s^{2*} .

21. A graph of the total test information function is obtained by `plot(TwoPL, type = 'info', theta_lim = c(-4,4))`. Figure 5.12 contains the corresponding plot. (We convert the standard blue color used for the function to black by `trellis.device(color = FALSE)`.)

22. On the normal metric the total information area index is

$$I_A = D \sum_{j=1}^L \alpha_j \iota \quad (5.26)$$

and item information area index (Birnbaum, 1968):

$$I_{A_j} = D \alpha_j \iota, \quad (5.27)$$

with $D = 1.702$.

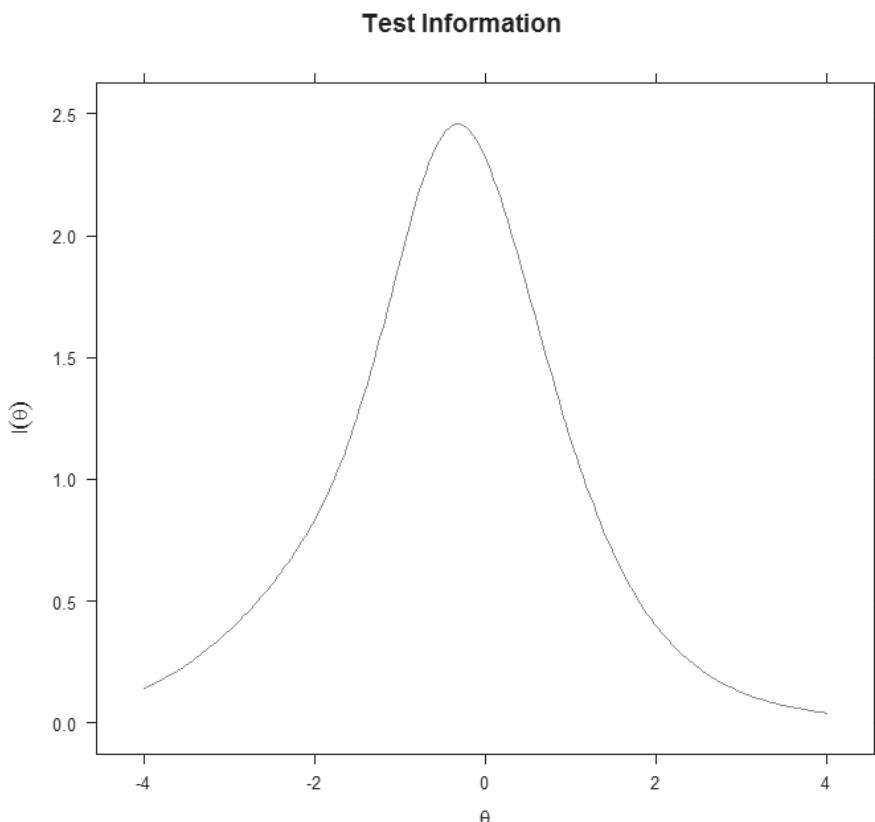


FIGURE 5.12. Total (test) information function.

23. These statements are tied to this particular metric. As a result, the 2PL and 1PL models' metrics were linked with one another prior to superimposing the two information functions. This is accomplished by using the total characteristic function equating approach to place the 1PL model estimates (i.e., the initial metric) on the 2PL model metric (i.e., the target metric); this approach is discussed in Chapter 11. The metric transformation coefficients are $\zeta = 1.0150$ and $\kappa = 0.0037$. These values indicate that the two metrics are in very close alignment even before linking, because values of $\zeta = 1$ and $\kappa = 0$ would indicate perfect alignment between the metrics. After linking, the 1PL estimates become $\hat{\alpha} = 1.40$, $\hat{\delta}_1 = -1.950$, $\hat{\delta}_2 = -0.586$, $\hat{\delta}_3 = -0.264$, $\hat{\delta}_4 = 0.292$, and $\hat{\delta}_5 = 0.453$.
24. With the 2PL and 3PL models (see Chapter 7), we have an indeterminacy of not only the metric's location, but also the metric's scale. As was the case with the Rasch/1PL model, the adding/subtracting of a constant, c_1 , to θ and δ simply shifts the metric up or down the continuum. With respect to the metric's scale, multiplying/dividing by a constant, c_2 , simply stretches/contracts the metric's scale. Moreover, if we divide α_j by c_2 and multiply the $(\theta - \delta_j)$ by c_2 we will obtain the same probability. That is, let $\alpha_j^* = \alpha_j/c_2$, $\theta_i^* = \theta_i + c_1$, and $\delta_j^* = \delta_j + c_1$, then our logistic deviate is $\alpha_j^*(\theta_i^* - \delta_j^*)$. Thus, $\alpha_j^*(\theta_i^* - \delta_j^*) = (\alpha_j/c_2)[(\theta_i + c_1) - (\delta_j + c_1)] = (\alpha_j/\delta_j)(\theta_i - \delta_j)$, so that when we multiply by c_2 , we obtain $c_2[(\alpha_j/c_2)(\theta_i - \delta_j)] = \alpha_j(\theta_i - \delta_j)$. As such, p_j has the same value on the transformed metric as on the untransformed metric. Stated another way, there is nothing unique about our estimates' metric in the sense that there are an infinite number of metrics that will provide the same answer. As mentioned above, we need to fix the metric's location by, for example item centering, in order to address this issue and identify the model. In addition, with the 2PL and 3PL models, we also need to fix the metric's scale by, for example, setting the variance to 1. Although in the foregoing we divide α_j by c_2 , we could have multiplied α_j by c_2 . In this case, with $\alpha_j^* = \alpha_jc_2$ we divide $(\theta - \delta_j)$ by c_2 .

6

The Three-Parameter Model

In this chapter we present a model for addressing chance success on an item. This chance success is reflected in an IRF with a nonzero lower asymptote. To model this lower asymptote, we extend the 2PL model to produce the three-parameter model. Parallel to the structure of the chapters discussing the 1PL and 2PL models, we present examples of a three-parameter model calibration using the mathematics data set introduced in Chapter 2.

Through the previous chapters we have developed a “toolbox” of model-fit techniques. This toolbox includes methods for assessing the tenability of various assumptions. To summarize these approaches, the unidimensionality assumption can be assessed using nonlinear factor analysis, linear factor analysis, and structural equation modeling. We can assess the tenability of the functional form assumption by examining the empirical IRFs. Moreover, model–data fit can be assessed through fit statistics (e.g., INFIT, OUTFIT, M2), comparing the predicted and empirical IRFs, as well as by obtaining evidence of item parameter estimate invariance through the use of several statistics (e.g., correlations, RMSD, UA₂₂). We have also examined person fit through fit statistics.

In this chapter we add to our toolbox. Specifically, (1) we introduce the likelihood ratio, AIC, and BIC statistics for making model comparisons, (2) we use Q_3 for assessing the tenability of the conditional independence assumption, and (3) we discuss the appropriateness of a person’s estimated location as a measure of their true location. Although for pedagogical reasons we present the model-fit techniques separately, in practice they would be used collectively. The last topic we cover in this chapter is the handling of missing data.

Conceptual Development of the Three-Parameter Model

Individuals at the lower end of the latent continuum may be expected to have a high probability of providing a response of 0. For example, examinees who have low mathematics proficiency may be expected to incorrectly respond to, say, a topology question

on a mathematics examination. If this mathematics examination uses a multiple-choice item format, then some of these low-proficiency individuals may select the correct option simply by guessing. Similarly, people low in neuroticism who are administered a neuroticism inventory using a true/false response format may be expected to respond “False” to a question depicting a neurotic behavior. However, owing to inattention or fatigue, some of these individuals may respond “True” to the question. In these cases, the item’s response function has a lower asymptote that may not be asymptotic with 0.0 but may be with some nonzero value. The three-parameter model addresses this non-zero lower asymptote.

To develop the three-parameter model, we need to be concerned with two cases. The first case is, “What is the probability of a response of 1 on an item when an individual responds consistent with their location θ ?” Our answer is that the probability of a response of 1 is modeled by the 2PL model. Conversely, the probability of a response of 0 (i.e., $p(x_j = 0 | \theta, \delta)$) when an individual responds consistent with their location θ is given by $(1 - p_j)$; Figure 2.12 depicts these two functions. The $p(x_j = 0 | \theta, \delta)$ response function has a lower asymptote of 1 and an upper asymptote of 0. That is, as θ approaches $-\infty$, the event “a response of 0” is almost certain to occur.

The second case to consider is, “What should be the probability of a response of 1 on an item due to chance alone?” To answer this question, let us symbolize this probability as χ_j . In other words, when a person can be successful on item j regardless of the person’s location, then the corresponding probability is given by χ_j . To determine the pseudo-random guessing response function, we need to consider χ_j and the probability of a response of 0 given the 2PL model (i.e., $p(x_j = 0 | \theta, \delta) = [1 - p_j]$). Noting that the event “a response of 1 due to chance alone” is independent of the event “a response of 0 given θ ” allows us to apply the multiplication rule. That is, when a person can be successful on item j on the basis of chance alone the probability is given by the pseudo-random guessing response, $\chi_j[1 - p_j]$. Multiplying by $[1 - p_j]$ transforms the lower asymptote of $p(x_j = 0 | \theta, \delta)$ to equal χ_j . Thus, as θ goes to $-\infty$, p_j approaches 0.0 and $\chi_j[1 - p_j]$ simplifies to χ_j . Conversely, as θ goes to ∞ , p_j approaches 1.0 and $\chi_j[1 - p_j]$ approaches 0.0. Thus, the probability of a response of 1 for an individual with an infinitely low location is χ_j .

Putting these two (mutually exclusive) cases together, we obtain the probability of a response of 1

$$p_j^* = p_j + \chi_j(1 - p_j), \quad (6.1)$$

where p_j is given by the 2PL model. Equation 6.1 may be rearranged to be

$$p_j^* = \chi_j + (1 - \chi_j)p_j. \quad (6.2)$$

By substitution of the 2PL model for p_j , we obtain the *three-parameter logistic* (3PL) model

$$p(x_j = 1 | \theta, \alpha_j, \delta_j, \chi_j) = \chi_j + (1 - \chi_j) \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}}. \quad (6.3)$$

One can view Equation 6.3 from a slightly different perspective than above and explain why Equation 6.2 is not simply $\chi_j + p_j$. The effect of the term $(1 - \chi_j)$ compresses the 2PL model's IRF to range from zero to $(1 - \chi_j)$. By adding χ_j to this compressed IRF (i.e., Equation 6.3), we transform the IRF to have a range from χ_j to 1.0. One implication of this compression is that it effectively reduces the IRF's slope.

Although, strictly speaking, Equation 6.3 is not in logistic form, it is referred to as a logistic model. (Because there is a normal ogive version of the three-parameter model, Equation 6.3 is sometimes presented, incorporating the scaling factor D .) As is the case with the 1PL and 2PL models, δ_j represents item j 's location and α_j reflects its discrimination parameter. The additional parameter, χ_j , is referred to as the item's *pseudo-guessing* or *pseudo-chance* parameter and equals the probability of a response of 1 when θ approaches $-\infty$ (i.e., $\chi_j = p(x_j = 0 | \theta, \rightarrow -\infty)$). As such, χ_j represents the IRF's lower bound or asymptote. With the 3PL model, there are three parameters characterizing the item j (i.e., $\alpha_j, \delta_j, \chi_j$) plus a person parameter.

The 3PL model is based on the same assumptions discussed in Chapter 2 with the 1PL model. Recall that these assumptions are a unidimensional latent space, conditional independence, and a specific functional form. For brevity we use p_j instead of $p(x_j = 1 | \theta, \alpha_j, \delta_j, \chi_j)$ in the following.

Examples of the 3PL model's IRF are given in Figure 6.1. The two items shown have the same discrimination and location parameters, but they have different χ_j s. For item 1 $\chi_1 = 0.1$ and for item 2 $\chi_2 = 0.05$. We see that the IRFs have nonzero lower asymptotes

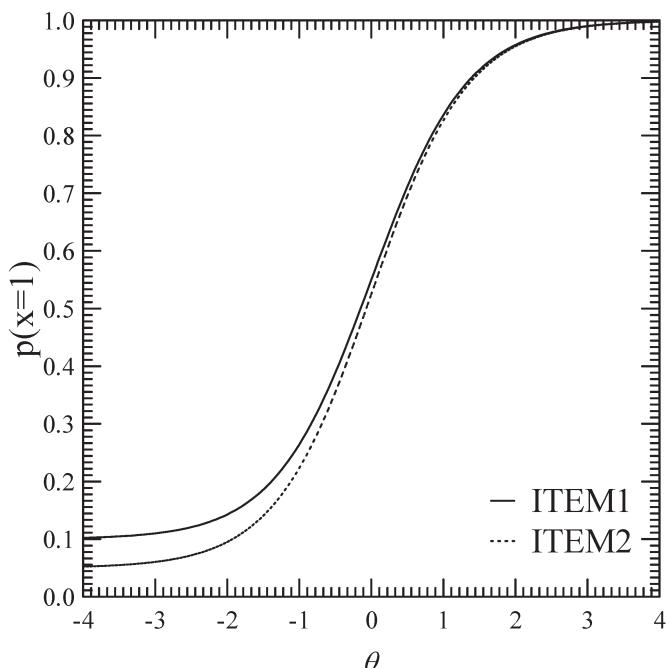


FIGURE 6.1. 3PL model IRFs for two items with $\alpha_1 = 1.5, \delta_1 = 0.0, \chi_1 = 0.1$, and $\alpha_2 = 1.5, \delta_2 = 0.0, \chi_2 = 0.05$.

and that each IRF is asymptotic with its corresponding χ_j value. In addition, we see that item 1 with the larger χ_j has the higher IRF. In general, as χ_j increases, so does p_j , all other things being equal. In the context of proficiency assessment, this means that items with larger χ_j s are easier than those with smaller χ_j s. The figure shows that the valid range for χ_j is 0.0 to 1.0.

As is the case with the 1PL and 2PL models, the IRF's slope is at a maximum at the item j 's location. This point of inflection occurs midway between the lower and upper asymptotes. The lower asymptote is the floor of the IRF and represents the smallest probability for a response of 1, whereas the upper asymptote is the ceiling for the IRF and reflects the largest probability of a response of 1. If we let Υ_j denote item j 's upper asymptote, then a general expression for determining the midpoint (i.e., the probability at δ_j) for any of our dichotomous models is $(\Upsilon_j + \chi_j)/2$. For example, with the 1PL and 2PL models, the lower asymptote is 0 and the upper asymptote is 1. Therefore, for the 1PL and 2PL models we have that $\chi_j = 0.0$, $\Upsilon_j = 1.0$, and the probability of a response of 1 at δ_j is $(1 + 0.0)/2 = 0.50$. For the 3PL model, if $\chi_j > 0.0$, then the probability of a response of 1 at δ_j is greater than 0.50. For example, if $\chi_j = 0.20$ and $\Upsilon_j = 1.0$, then the probability of a response of 1 at δ_j is $(1 + \chi_j)/2 = (1 + 0.2)/2 = 0.6$.¹ Moreover, as is true with the 1PL and 2PL models, the 3PL model's discrimination parameter is proportional to the slope at the inflection point. However, the relationship between α_j and the slope now involves χ_j . Specifically, the slope for the 3PL model is $0.25\alpha_j(1 - \chi_j)$.² Therefore, an item's discriminatory effectiveness is affected by the magnitude of χ_j . Specifically, as χ_j increases, an item's discriminatory effectiveness decreases, all other things being equal. For example, we see from Figure 6.1 that item 1's discriminatory effectiveness (reflected in its IRF's slope) is less than that of item 2.

Additional Comments about the Pseudo-Guessing Parameter, χ_j

Our first comment is about the different labels used for χ_j . Originally, χ_j was referred to as the item's guessing parameter (e.g., Lord, 1980, p. 12). However, because χ_j is typically lower than what would be predicted by a random guessing model (i.e., the reciprocal of the number of multiple-choice options), χ_j is now referred to as the pseudo-guessing parameter. This difference between χ_j and the random guessing model prediction is due to differential option attractiveness. That is, the random guessing model assumes that all options are equally attractive. Yet we know from traditional item analyses that item alternatives vary in their degree of attractiveness to persons. For instance, using keywords in alternatives is a typical tactic to increase the attractiveness of alternatives. Moreover, test-taking preparation instructs examinees who do not know the answer to a question to select the longest option because it is usually the correct response. As such, the random guessing model's assumption is not reflected in the response data.

Our second comment concerns the nature of χ_j . As mentioned earlier, χ_j 's function is to reflect that some individuals with infinitely low locations may obtain a response of 1 when, according to the 2PL model, they should not. These responses are a manifesta-

tion of the interaction between person and item characteristics (including item format). In the case of proficiency instruments, person characteristics include not only a person's θ , but also their test-wiseness and "risk-taking" tendencies. These last two factors are tangential latent person variables. Therefore, although χ_j is considered to be an item parameter, it may be more reflective of a person characteristic (i.e., another person parameter) than of an item characteristic or, at least, an interaction between person and item characteristics.

Our final comments concern the implicit assumption made by the use of χ_j and the effect of χ_j on estimation. In regard to the former, we see from Equation 6.3 that the presence of χ_j in the model assumes that, regardless of a person's location, their propensity to "guess" is constant across the continuum (i.e., χ_j does not vary as a function of θ). This assumption may or may not be reasonable in all situations. With respect to effects, nonzero χ_j 's lower the estimate of a person's location (Wainer, 1983) and reduce the amount of item information.³ Thus, although we are modeling nonzero χ_j 's, it is very desirable that our χ_j 's be close to zero. Of course, in this case the 2PL model may provide a sufficiently reasonable representation of the data.

Conceptual Parameter Estimation for the 3PL Model

The estimation of item parameters proceeds as discussed in previous chapters. However, unlike the 1PL and 2PL models, the 3PL model does not have sufficient statistics for parameter estimation (Baker, 1992; Lord, 1980). The log likelihood surface for an item with three item parameters would require four dimensions to graphically represent it. However, the general idea can be represented as a series of static multiple surfaces similar to the one presented in Figure 5.3, but with each surface slightly different from the others and associated with a particular value of χ_j (e.g., 0.0, 0.01, 0.02). (Obviously, the discrete nature of this series of surfaces does not accurately reflect the continuous nature of χ_j .) The essence of the estimation process would be to identify across these "multiple surfaces" the values of α_j , δ_j , and χ_j that maximize the log likelihood for an item.⁴

In some cases, distinguishing between these multiple surfaces may be problematic. For instance, if there are insufficient data at the lower end of the continuum, then there may be multiple sets of α_j , δ_j , and χ_j that account for the data. As such, the corresponding IRFs are similar to one another in this region (cf. Mislevy, 1986a). As an example, assume that in a given calibration sample everyone is located above -1. As a result, there is insufficient data to estimate the lower asymptote. Figure 6.2 presents two IRFs that can account for empirical data. One IRF is based on $\alpha = 0.8$, $\delta = -0.05$, and $\chi = 0.435$, whereas the other has the item parameter values of $\alpha = 0.56$, $\delta = -1.8$, and $\chi = 0.0$. As can be seen, these two IRFs are very similar to one another above -1 and, in fact, differ by less than 0.01 in the θ range -1 to 1 and by less than 0.018 in the range -1 to 3. Without additional information (e.g., persons located around -3, or prior information), it is not possible to determine whether χ_j should be 0.435 or 0. In terms of our "multiple surfaces" analogy, this means that we cannot distinguish between the log likelihood

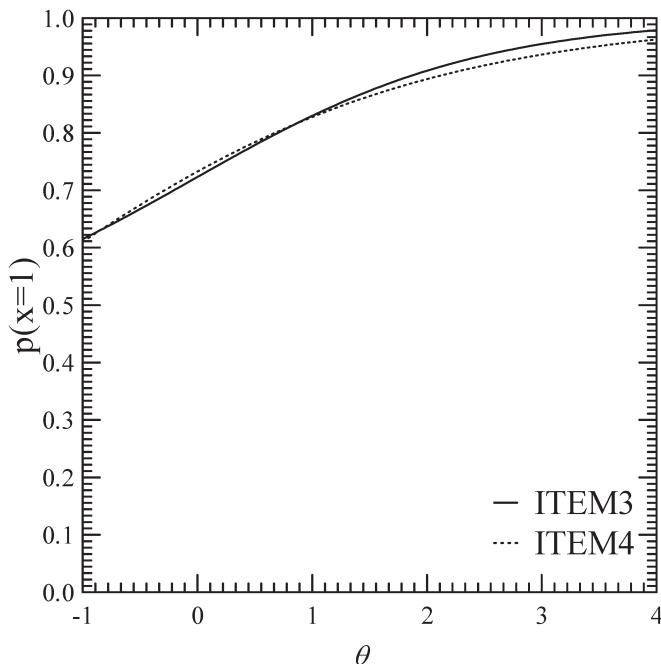


FIGURE 6.2. 3PL model IRFs when $\alpha_1 = 0.8$, $\delta_1 = -0.05$, $\chi_1 = 0.435$ and when $\alpha_2 = 0.56$, $\delta_2 = -1.8$, $\chi_2 = 0.0$.

surface associated with $\chi = 0.435$ and the one when $\chi = 0.0$. Therefore, if the respondents are located above -1 , it is difficult to determine which of these two sets of item parameter estimates is “best,” and so we have difficulty obtaining a converged solution for the item.⁵

In general, the estimation of χ_j may be problematic for some items because of the paucity of persons at the lower end of the continuum; because the items are located at the lower end of the continuum (e.g., very easy items); and/or because the items have low estimated discrimination parameters. Problems in estimating χ_j can influence the estimation of the item’s other parameters. In these situations, a criterion may be used to determine whether χ_j should be estimated. For instance, LOGIST used the “stability” criterion of $(\delta_j - 2/\alpha_j)$. Specifically, χ_j is estimated only when $(\delta_j - 2/\alpha_j) > -2.5$; -2.5 is the default value and may be changed. The stability criterion is the location on the θ continuum “at which the proportion of correct responses is only about 0.03 above the lower asymptote” (Wingersky et al., 1982, p. 21). Alternative strategies are to fix χ_j to a specific value or to impose a prior distribution. With respect to the former, the selection of a constant (common) value for χ may be done arbitrarily (e.g., LOGIST’s $[1/m - 0.05]$ where m is the number of item options), by averaging the nonproblematic $\hat{\chi}_j$ s, by averaging the $\hat{\chi}_j$ s for items located at the lower end of the continuum, or by fixing the lower asymptote to some nonzero value determined by inspecting the lower asymptote of empirical IRFs.

We may also use a prior distribution with χ_j . de Gruijter (1984) has demonstrated that the use of a prior distribution for estimation of χ_j can lead to reasonable parameter estimates for the model. The regression toward the mean phenomenon that typically occurs when using a prior distribution is not as problematic in estimating χ_j as it is when estimating person and item location parameters (Lord, 1986). In general, we recommend use of a prior on the $\hat{\chi}_j$ s as the first strategy to facilitate estimating the lower asymptote.

In regard to the item's other parameters, empirical data calibration has shown that the $\hat{\alpha}_j$ s and $\hat{\delta}_j$ s are nonlinearly related and, typically, have a positive correlation (Lord, 1975). In addition, Lord found that items with $\hat{\delta}_j$ s less than about -0.5 almost never have $\hat{\alpha}_j$ s greater than 1 and that items located above 0.5 almost always have $\hat{\alpha}_j$ s greater than 1.0. In this regard, we examined the calibration results from the reading and mathematics tests from the National Education Longitudinal Study, 1988 (NELS: 88; Ingels, Scott, Rock, Pollack, & Rasinski, 1994) base year, and found the correlation between the $\hat{\alpha}_j$ s and the $\hat{\delta}_j$ s is 0.25 for the reading test and 0.59 for the mathematics test, the 3PL model calibration used LOGIST; also see Yen (1987). Baker and Kim (2004) present the mathematics for estimating the three item parameters, and a Bayesian estimation procedure is presented in Swaminathan and Gifford (1986).

So far we have been concerned with item parameter estimation. We now turn our attention to person parameter estimation. Any of the methods that were previously discussed, such as MLE or EAP, could be used. However, in some cases the use of unrestricted MLE for person location estimation may encounter problems. For example, Samejima (1973a) showed that there is not a unique solution for θ for every possible response pattern under the three-parameter model. For these problematic response patterns, the likelihood function may have more than one maximum. For example, assume we have a two-item instrument with $\alpha_1 = 2.0$, $\delta_1 = 0.0$, $\chi_1 = 0.25$ for the first item and $\alpha_2 = 1.0$, $\delta_2 = -0.5$, $\chi_2 = 0.0$ for the second item. On these two items, assume that a person has a response of 1 on item 1 and a response of 0 on item 2 (Samejima, 1973a). Assuming a proficiency testing situation, then this response pattern reflects a person correctly answering the "harder/more discriminating" item (possibly by guessing) and incorrectly answering the "easier/less discriminating" item. The corresponding likelihood function is presented in Figure 6.3.

As we see, the likelihood function has a *local* maximum at approximately -0.05, and as θ becomes progressively smaller, the likelihood function begins to approach an asymptote of 0.25. Therefore, this likelihood function is asymptotic with a value of 0.25 and without a unique person location estimate. Stated another way, these types of response vectors do not have a *global* maximum and have multiple maxima. In these cases, the use of standard MLE with the three-parameter model may yield a $\hat{\theta}$ that turns out to be a local, not a global, maximum. When a local maximum is suspected, then using a different starting/provisional estimate for the MLE algorithm (see Appendix A) might produce a different $\hat{\theta}$. (In fact, the presence of multiple solutions for a given response vector is evidence that one or more of the solutions represent local maxima.)

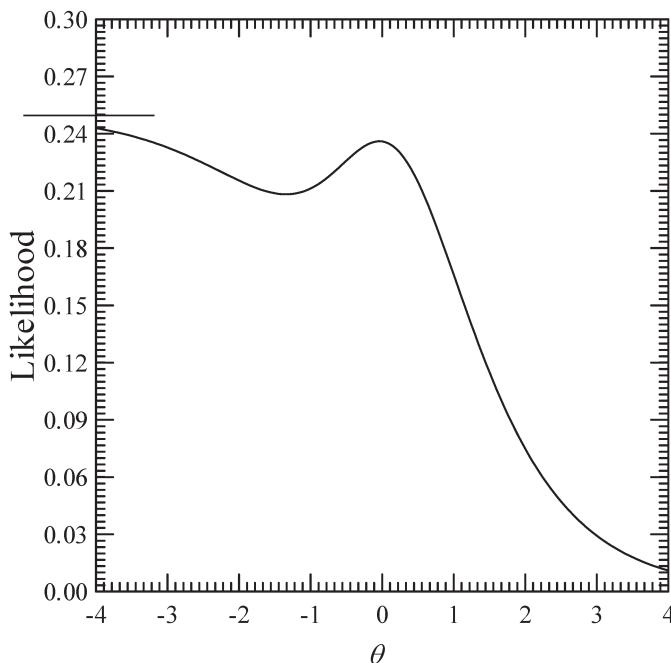


FIGURE 6.3. Likelihood function for a two-item instrument with no unique maximum in which the first response is correct and the second is incorrect (i.e., $\bar{x}' = 10$).

Although the example only uses two items, Samejima (1973a) speculated that “the likelihood function may be more complicated, with possibly more than one local maximum in addition to a terminal maximum” (p. 225). We have empirical support for the occurrence of multimodal likelihood functions from the work of Yen, Burkett, and Sykes (1991). Specifically, in an analysis of 14 empirical data sets they found that as many as 3.1% of the examinees had response vectors whose likelihood functions had multiple maxima (cf. Fischer, 1981).

The multimodal likelihood function seen in Figure 6.3 is due to the specific \bar{x} and the particular relationship among the α_j , δ_j , and χ_j . If $\chi_j = 0$ for both items (i.e., the 1PL and 2PL models), then the likelihood function has a unique solution. Therefore, one possibility of addressing these multimodal likelihood function situations is to use the *truncated 2PL model* (Samejima, 2001) for person parameter estimation. The truncated 2PL model capitalizes on the fact that the 2PL model’s IRF (with appropriate item parameter values) is virtually indistinguishable from that of the 3PL model above a critical value, θ_g . Below θ_g the probability of a response of 1 is 0 for the truncated 2PL model (i.e., the IRF is truncated at θ_g). Therefore, for the truncated 2PL model there are two conditions: (1) for $-\infty < \theta < \theta_g$ where we have that $p_j = 0$; and (2) for $\theta_g < \theta < \infty$; p_j is given by the 2PL model (Equation 5.1). Samejima (1973a) shows that $\theta_g = 0.5 \ln(\chi_j) + \delta_j$. An alternative approach for handling multimodal likelihood functions is to use a Bayesian person estimation technique (e.g., EAP).

How Large a Calibration Sample?

The answer to the question of how large a sample is needed depends, in part, on the estimation procedure, instrument characteristics (e.g., the distribution of item parameter estimates, instrument length, etc.), response data characteristics (e.g., amount of missing data), and person distribution. In general, attempts at answering this question have involved conducting simulation studies where the parameter estimates can be compared, directly or indirectly, with the corresponding parameters.

For example, Yen (1987) investigated the parameter recovery of MMLE and JMLE as implemented in BILOG and LOGIST, respectively, using a fixed sample size of 1,000. In this Monte Carlo study, she investigated three different instrument lengths (10, 20, and 40 items) and different θ distributions: normal (0, 1), negatively skewed ($\text{skew} = -0.4/\text{kurtosis} = -0.1$), positively skewed ($\text{skew} = 0.4/\text{kurtosis} = -0.1$), and platykurtic ($\text{skew} = 0.1/\text{kurtosis} = -0.4$). Generally speaking, she found that MMLE estimates were more accurate than those of JMLE, particularly at the 10-item instrument length. With respect to MMLE, the item discrimination estimation results using the 20- and 40-item instruments were comparable to one another in terms of their RMSD, with values ranging from 0.09 to 0.20. Moreover, the correlations between $\hat{\alpha}$ and α ($r_{\alpha\hat{\alpha}}$) ranged from 0.88 to 0.94, regardless of the normality or non-normality of the θ distribution. For the 10-item length, the RMSD doubled to 0.48 and $r_{\alpha\hat{\alpha}}$ decreased to 0.84. In terms of estimating item locations, the 20- and 40-item instruments had correlations ($r_{\delta\hat{\delta}}$) from 0.97 to 0.99 with RMSDs of 0.07 to 0.16; the 10-item length had an $r_{\delta\hat{\delta}}$ of 1.00 and RMSDs of 0.18. In general, item location is better estimated than item discrimination. The lower asymptote showed $r_{\chi\hat{\chi}}$ s between 0.11 and 0.54, with RMSDs of 0.03 to 0.08 across the various instrument lengths and irrespective of the nature of the θ distribution. Although not a formal parameter recovery study, Mislevy (1986a) presents results indicating that BILOG does a reasonably good job in recovering item parameters with a sample size of 1,000 and a 20-item instrument.

This research appears to indicate that for MMLE a sample of 1,000 persons may lead to reasonably accurate item parameter estimates with the 3PL model under favorable conditions (e.g., a symmetric θ distribution, an instrument length of 20 items). This rough guideline assumes the use of prior distributions for χ_j and α_j . However, it is strongly recommended that calibration sample sizes exceed 1,000 to mitigate the convergence problems that sometimes plague 3PL model calibrations. In fact, Thissen and Wainer (1982) suggest trying to avoid estimating χ_j if possible under unrestricted MLE, and they also suggest that the use of a prior distribution when estimating χ_j seems “to offer some hope” (p. 410). In cases where one has a smallish sample size and/or one experiences difficulty in estimating the item parameters with the 3PL model, then fixing the lower asymptote to a reasonable nonzero value for some or all the items may help. In addition, some convergence problems (e.g., $-2\ln L$ values that oscillate across iterations) may sometimes be rectified by using the RIDGE subcommand available in BILOG and PARSCALE. The calibration sample size caveats and considerations previously mentioned in Chapters 3 and 5, such as model–data misfit tolerance, ancillary

technique sample size requirements, the amount of missing data, and so on are also applicable to the three-parameter model.⁶

Assessing Conditional Independence

In Chapter 2, we stated that one assumption underlying IRT models is that the responses to one item are not related to those on any other item conditional on $\theta(s)$. This assumption is the conditional (or local) independence assumption. When this assumption is violated, then the accuracy of our item parameter estimates is affected and the total instrument information is overestimated (Chen & Thissen, 1997; Oshima, 1994; Sireci, Wainer, & Thissen, 1991; Thissen, Steinberg, & Mooney, 1989; Yen, 1993). As such, any subsequent use of the item parameter estimates for, say, equating (see Chapter 11) will be potentially adversely affected. In the following, we discuss some causes of item dependence, some ways to handle this dependence, and then a statistic for identifying local dependent items post administration.

Violation of the conditional independence assumption may occur for various reasons, such as structural dependence among items, content clues, instrument length, insufficient allotted time to complete an instrument (i.e., speededness), and/or an insufficient number of latent variables in the IRT model. Examples of items with structural dependence are a set of survey questions that all refer to the same, say, life-changing event (e.g., a diagnosis of cancer, contracting HIV), comprehension questions that use the same reading passage, or trigonometry problems based on a common figure. In all of these cases, one may see local dependence. In addition, when there is insufficient time to respond to all the items on an instrument, the items affected by the lack of time may exhibit dependence. As a consequence, their corresponding parameter estimates are adversely affected. Conversely, when there is sufficient time to respond to an instrument but the instrument is very long, one may observe local dependence due to fatigue or diminished motivation. Practice effects may also lead to local dependence.

For some of these causes, it is possible to identify the items that may be prone to local dependence prior to administering the instrument. In general, an instrument should be inspected for connections between the items. This inspection involves looking for similarity in the questions' text, an item providing one or more cues as to how to respond to another item, the items sharing grammatical inconsistencies or common information (e.g., a passage or a figure), the items sharing a nesting/hierarchical relationship, and so on. Depending on the outcome of this inspection, rewriting the items may be sufficient to address the anticipated dependency. In other cases, the items cannot be rewritten because they need to be logically related or structurally dependent. In these cases the dependent items may be combined to form an item cluster.

An item cluster (also known as an item bundle or testlet [Thissen, Steinberg, & Mooney, 1989b; Wainer & Kiely, 1987; Wainer & Lewis, 1990]) is a group of interdependent items that may be created pre- or post administration. There are at least two ways to score an item cluster. In one approach, each item in the item cluster provides an item score and the score on the item cluster is, for example, the sum of these item

scores. For instance, if an item cluster consists of three 1-point items, then the possible scores on the item cluster would be 0, 1, 2, or 3. In effect, the item cluster is treated as a single “item” for estimating a person’s location. One way of utilizing this item cluster score is to use a model that can handle both dichotomous and polytomous responses (e.g., see Yen, 1993). Models that can address not only polytomous responses, but also dichotomous responses, are presented in the following chapters.

In the foregoing polytomous model approach to handling item clusters, there is some loss of information. For example, an item cluster score does not say anything about the response pattern that produced the score. Whether this is an important issue is context-specific. However, if the loss of this information is important, then an alternative approach to scoring an item cluster is to use a model that incorporates a parameter that reflects the dependency among items within the item cluster. Bradlow, Wainer, and Wang (1999) developed such a model by augmenting the 2PL model. The augmentation is a random effect parameter that reflects a person-specific *testlet* effect.⁷ The Bradlow et al. model may be applied to both items that are independent and those in testlets; one- and three-parameter models also exist (see Wang & Wilson, 2005; Wainer, Bradlow, & Du, 2000). This *testlet model* and its variants form the basis of testlet response theory (Wainer, Bradlow, & Wang, 2007).

Various indices have been developed for identifying local dependence. A review of some of these indices and an examination of which index works best may be found in Kim, de Ayala, Ferdous, and Nering (2011); also see Glas (1999), Glas and Falcón (2003), Orlando and Thissen (2000), and Rosenbaum (1984) for related indices. One of these indices is Yen’s (1984) Q_3 index. Although no index may be considered to be the best in terms of combining high power to detect conditional item dependence with low false positive rates, the Q_3 index works reasonably well (e.g., see Kim et al., 2011). Because of Q_3 ’s simplicity and its comparative good performance, we use it to demonstrate evaluating the conditional independence assumption with our mathematics data example.

Q_3 is the correlation between the residuals for a pair of items. The residual for an item is the difference between an individual’s observed response and their expected item response. Therefore, after fitting the model, the Pearson correlation coefficient is used to examine the linear relationship between pairs of residuals. In the current context, the observed response is either a 1 or a 0 and the expected response is the probability according to the 3PL model. Symbolically, the residual for person i on item j is

$$d_{ij} = x_{ij} - p_j(\hat{\theta}_i)$$

and for item k it is

$$d_{ik} = x_{ik} - p_k(\hat{\theta}_i).$$

Q_3 is the correlation between d_{ij} and d_{ik} across respondents

$$Q_{3(j,k)} = r_{d_{ij}d_{ik}} \quad (6.4)$$

If $|Q_3|$ equals 1.0, then the two items are perfectly interdependent. In contrast, a Q_3

of 0.0 is a necessary, but not sufficient, condition for independence because a $Q_3 = 0$ can be obtained when the items in the pair are independent of one another or because they exhibit a nonlinear relationship. Therefore, Q_3 is useful for identifying items that exhibit item dependence. Under conditional independence Q_3 should have an expected value of $-\frac{1}{L(L-1)}$ (Yen, 1993).

As has been mentioned, in some cases one can explain item dependence in terms of multidimensionality. That is, the dependency between two items is due to a common additional latent variable such as test-wiseness. If two items are independent, then their interrelationship is completely explained by the latent structure of the model. If one applies a unidimensional model when two dimensions are called for, then the items that are influenced by both latent variables show a negative local dependence, and items that are affected by only one of the two latent variables show a positive local dependence (Yen, 1984). However, if only one of the latent variables is used, then the items that are influenced only by that underlying variable show a slight negative local dependence. To obtain a large Q_3 value for an item pair, we need to have similarity of parameters for the items in question and the items need to share one or more unique dimensions. Therefore, similarity of parameters is a necessary, but not sufficient, condition for obtaining a large Q_3 value.

Some research (e.g., Yen, 1984) has found that although the value of Q_3 is not as much influenced by the sample size as other measures, it is affected by the instrument's length. This may be due to item scores being involved in both x_{ij} and x_{ik} as well as (implicitly) in $p_j(\hat{\theta})$. As a result, Q_3 may tend to be slightly negative due to part-whole contamination (Yen, 1984). The implication is that one would expect to see substantially more negative Q_3 s for short instruments than for longer instruments. In this case, these negative values may be artifacts due to the instrument's length.

There are various ways to use Q_3 to identify locally dependent items. First, we can use Q_3 in a statistical z -test (Yen, 1984). This would require that Q_3 be transformed by the Fisher r -to- z transformation (\dot{z}_{Q_3}) and then used in a z -test,

$$z = \frac{\dot{z}_{Q_3}}{\sqrt{1/(N-3)}}$$

\dot{z}_{Q_3} has a mean of 0.0 and a variance of $1/(N - 3)$. The standard unit normal table is used to provide critical values for identifying items with \dot{z}_{Q_3} values that are unlikely to be observed owing to chance alone. However, because the typical calibration sample size will result in a test with a great deal of power, we will most likely reject the null hypothesis of independence for trivially small correlations. An additional issue is that the sampling distribution of Q_3 may not be symmetric (Chen & Thissen, 1997). That is, because the Q_3 sampling distribution may not approximate the standard normal very well, the critical values would be inappropriate. As a result, the Q_3 empirical Type I error rates do not match the nominal significance level that one would expect under normal curve theory. Moreover, Marais (2013) states that the "sampling properties of the correlations among residuals are unknown. It is therefore not possible to use these statistics for for-

mal tests of local independence" (p. 121). Therefore, rather than using the critical values from the standard unit normal table in a statistical inferential fashion, it is preferable to use them as guidelines/screening values for informed judgment.

A second way of using Q_3 is to take advantage of the fact that Q_3 is a correlation. Specifically, because Q_3 is a correlation coefficient, the square of Q_3 (Q_3^2) may be interpreted as a measure of the amount of residual variance shared by an item pair. Therefore, item pairs with a large proportion (e.g., 5% or greater) of shared variability would indicate dependent items.

Alternatively, one could compare Q_3 to a cutpoint. That is, an observed Q_3 that is larger than the cutpoint would indicate item dependence. Yen (1993) suggests one such cutpoint in the context of instruments with a minimum of 17 items. Specifically, a Q_3 screening value of 0.2 was suggested to identify items exhibiting dependence (i.e., $|Q_3| > 0.2$ indicates local item dependence). Although this cutpoint has been found to produce small Type I error rates, it also leads to comparatively lower power than other detection methods (Chen & Thissen, 1997).

To address some of these issues, Christensen, Makransky, and Horton (2017) conducted a study to arrive at empirically based critical values. Using simulation in conjunction with empirical estimates, they found, for example, that critical values of 0.19 and 0.24 cut off 5% and 1%, respectively, of $\max(Q_{3(j,k)})$'s empirical distribution above them with a 9-item instrument; $\max(Q_{3(j,k)})$ is the largest observed $Q_{3(j,k)}$. Although the study focused on the Rasch model, their results may be considered indicative that no single critical value can be used in all situations regardless of the IRT model used. They concur with Marais's (2013) conclusion that Q_3 's evaluation should take into account all of an instrument's Q_3 s (cf. Marais, 2013, p. 121). Specifically, a given $Q_{3(j,k)}$ is compared to " $\max(Q_{3(j,k)}) - \bar{Q}_3$ ", where Christensen et al. (2017) define the average Q_3 as

$$\bar{Q}_3 = 2 \sum_{j>k} Q_{3_{jk}} / (L(L-1)).$$

As Equation 6.4 shows, Q_3 is the zero-order correlation for item j 's and k 's respective residuals. (In this paragraph, all references to items are to the items' residuals.) As such, unless all item pairs are independent of one another, the correlation between item j and k will contain information from the other items to varying degrees. For example, assume that we have a three-item instrument and the correlation between items 1 and 2 is -0.181 . If the correlation between items 1 and 3 is zero and the correlation between items 2 and 3 is also zero, then the correlation between items 1 and 2 is the zero-order correlation (e.g., $Q_3 = r_{d_1 d_2} = -0.181$). However, if the correlation between items 1 and 3 is -0.098 and if the correlation between items 2 and 3 is -0.304 , then our zero-order correlation's magnitude is affected by the linear relationships item 3 has with items 1 and 2. To obtain an accurate assessment of the linear relationship between items 1 and 2, we should remove the linear influences of item 3 on items 1 and 2. Thus, we introduce a modified Q_3 statistic in which we calculate the $(L-2)$ -order partial correlation for items j and k

$$Q_3^P = r_{d_j d_k \cdot z}, \quad (6.5)$$

where \bar{z} represents all the instrument's items except items j and k . For our example, the first-order partial correlation between items 1 and 2 removing the linear effects of item 3 is $Q_3^P = -0.222$. Comparing Q_3^P and Q_3 shows item 3's linear effect on Q_3 . In short, in this case Q_3 shows less item dependency between items 1 and 2 than does Q_3^P . As is the case with Q_3 , large values of Q_3^P reflect item dependence, with values around 0 indicating either no linear relationship between items j and k or item independence.

Example: Application of the 3PL Model to the Mathematics Data, MMLE, BILOG-MG

A number of programs perform 3PL model calibration, including, but not limited to, BILOG-MG, XCALIBRE, mirt, NOHARM, SAS proc irt, and TAM. For comparison with our previous calibrations we use BILOG and then mirt.

Table 6.1 shows the command file for performing the calibration. As can be seen, we estimate both item and person parameters in a single run and save both the item estimates (PARm = 'MATH3PL.PAR') and person location estimates (SCOrE = 'MATH3PL_EAP.SCO') using the SAVE subcommand on the GLOBAL command line and the SAVE command.

Table 6.2 contains the corresponding abridged Phases 1 and 2 output. The echo of the program parameters indicates that the intended model (3 PARAMETER LOGISTIC) and the logistic metric (LOGIT METRIC) are being used. The echo of the Phase 2 program parameters shows the maxima of 50 EM and 20 Newton cycles (CALIB line) as well as the default convergence criterion of 0.01. Moreover, the output indicates the use of prior distributions for the estimation of α_j and χ_j (i.e., CONSTRAINT DISTRIBUTION ON SLOPES and CONSTRAINT DISTRIBUTION ON ASYMPTOTES, respectively).

The Phase 2 results show convergence in 14 EM cycles, and 3 Newton cycles were executed. The item parameter estimates from the converged solution are presented in Table 6.3. The item discrimination, location parameter, and pseudo-guessing parameter estimates are obtained from the SLOPE, THRESHOLD, and ASYMPtote columns,

TABLE 6.1. BILOG Command File for the 3PL Model Item Calibration

```
Example 3PL Calibration w/ person scoring

>GLOBAL DFName = 'C:\Math.dat', NPArm = 3, LOGistic, SAVE;
>SAVE PARm = 'MATH3PL.PAR', SCOrE = 'MATH3PL_EAP.SCO';
>LENGTH NITems = (5);
>INPUT NTotAl = 5, NIDchar = 10;
>ITEMS ;
>TEST1 TNAmE = 'TEST0001', INUmber = (1(1)5);
(10A1, T1, 5(1X,A1))
>CALIB CYCLES=50, NEWTON=20, PLOT = 1.0000, ACCel = 1.0000,
   CHIsquare = (5, 8);
>SCORE ;
```

TABLE 6.2. BILOG Output: Phases 1 and 2 (Abridged)

<Phase 1 results>

```

>GLOBAL DFNAME='MATHPAT.DAT', NPARM=3, NWGHT=3, LOG, SAVE;

FILE ASSIGNMENT AND DISPOSITION
=====
SUBJECT DATA INPUT FILE      C:\MATH.DAT
BILOG-MG MASTER DATA FILE    MF.DAT
                                         WILL BE CREATED FROM DATA FILE

CALIBRATION DATA FILE        CF.DAT
                                         WILL BE CREATED FROM DATA FILE

ITEM PARAMETERS FILE          IF.DAT
                                         WILL BE CREATED THIS RUN

CASE SCALE-SCORE FILE        SF.DAT
CASE WEIGHTING                NONE EMPLOYED

ITEM RESPONSE MODEL           3 PARAMETER LOGISTIC
                               LOGIT METRIC (I.E., D = 1.0)
19601 OBSERVATIONS READ FROM FILE:  C:\MATH.DAT
19601 OBSERVATIONS WRITTEN TO FILE: MF.DAT
                                         ::

ITEM STATISTICS FOR SUBTEST TEST0001

          ITEM*TEST CORRELATION
ITEM   NAME     #TRIED   #RIGHT   PCT    LOGIT    PEARSON   BISERIAL
-----+
 1  ITEM0001  19601.0  17395.0  88.7  -2.07    0.246    0.407
 2  ITEM0002  19601.0  12624.0  64.4  -0.59    0.439    0.564
 3  ITEM0003  19601.0  11094.0  56.6  -0.27    0.416    0.524
 4  ITEM0004  19601.0   8369.0  42.7   0.29    0.405    0.511
 5  ITEM0005  19601.0   7592.0  38.7   0.46    0.312    0.397
-----+

```

<Phase 2 results begin>

```

CALIBRATION PARAMETERS
=====
MAXIMUM NUMBER OF EM CYCLES:      50
MAXIMUM NUMBER OF NEWTON CYCLES:  20
CONVERGENCE CRITERION:            0.0100
ACCELERATION CONSTANT:           1.0000
LATENT DISTRIBUTION:              NORMAL PRIOR FOR EACH GROUP
PLOT EMPIRICAL VS. FITTED ICC'S: YES, FOR ITEMS WITH FIT PROBABILITY
                                     LESS THAN 1.00000
DATA HANDLING:                   DATA ON SCRATCH FILE
CONSTRAINT DISTRIBUTION ON ASYMPTOTES: YES
CONSTRAINT DISTRIBUTION ON SLOPES:   YES
CONSTRAINT DISTRIBUTION ON THRESHOLDS: NO
SOURCE OF ITEM CONSTRAINT DISTIBUTION
MEANS AND STANDARD DEVIATIONS:    PROGRAM DEFAULTS
                                     ::

METHOD OF SOLUTION:
EM CYCLES (MAXIMUM OF      50)
FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF  20)
                                     ::


```

(continued)

TABLE 6.2. (continued)

```
[EM STEP]

-2 LOG LIKELIHOOD =      111772.062
CYCLE      1;    LARGEST CHANGE=   0.35458
-2 LOG LIKELIHOOD =      110088.005
:
-2 LOG LIKELIHOOD =      110066.473
CYCLE     14;    LARGEST CHANGE=   0.00782

[FULL NEWTON CYCLES]
-2 LOG LIKELIHOOD:      110065.8456
CYCLE     15;    LARGEST CHANGE=   0.10845
:
-2 LOG LIKELIHOOD:      110066.0225
CYCLE     17;    LARGEST CHANGE=   0.00327
:
```

respectively. For instance, for item 1 the item discrimination estimate ($\hat{\alpha}_1$) is 1.608, and the item location estimate ($\hat{\delta}_1$) is -1.561, with a pseudo-guessing parameter estimate ($\hat{\chi}_1$) of 0.228. By and large, the $\hat{\chi}_j$ s are acceptable.⁸

As part of our model-data fit analysis, we compare the empirical and predicted IRFs for each of our items. Figure 6.4 shows item 4's empirical and predicted IRFs. The estimate of the item's pseudo-guessing parameter is identified by the symbol c instead of

TABLE 6.3. BILOG Output: Phase 2 (Abridged)

ITEM	INTERCEPT		SLOPE	THRESHOLD	LOADING	ASYMPTOTE	CHISQ (PROB)	DF
	S.E.	S.E.	S.E.	S.E.	S.E.	S.E.		
ITEM0001	2.510	1.608	-1.561	0.849	0.228	693.1	(0.0000)	4.0
	0.160*	0.092*	0.154*	0.049*	0.096*			
ITEM0002	0.661	2.802	-0.236	0.942	0.156	826.5	(0.0000)	3.0
	0.100*	0.217*	0.050*	0.073*	0.029*			
ITEM0003	-0.328	2.397	0.137	0.923	0.202	665.4	(0.0000)	4.0
	0.121*	0.183*	0.041*	0.071*	0.020*			
ITEM0004	-1.482	2.788	0.532	0.941	0.147	495.5	(0.0000)	5.0
	0.177*	0.254*	0.022*	0.086*	0.011*			
ITEM0005	-1.420	1.608	0.883	0.849	0.156	611.5	(0.0000)	5.0
	0.153*	0.134*	0.033*	0.071*	0.016*			

* STANDARD ERROR

LARGEST CHANGE =	0.003266	3291.9	21.0
		(0.0000)	

:

χ and the item's location by b . As can be seen, there is a close correspondence between the empirical and predicted IRFs. This correspondence provides evidence of data fit for this item. Figure 6.4 is typical of the other empirical versus predicted IRFs plots. Item 4's IRF and information function are presented in the left and right panels, respectively, of Figure 6.5. From the right panel we see that the item's information function is not quite centered about the item's location. This is true for all items calibrated with the 3PL model and is due to the influence of a nonzero lower asymptote. The actual location of the maximum of the item information is discussed below.

Fit Assessment: Conditional Independence Assessment

We use Q_3^P for evaluating the conditional independence assumption; Appendix G "Conditional Independence Using Q_3 " shows the analysis using Q_3 and a simulation approach for identifying a screening value. With a five-item instrument, there are 10 Q_3^P values to calculate (i.e., $L(L - 1)/2$). To calculate Q_3^P we need to have person location estimates to calculate the expected responses, p_j s. Our (EAP) estimates are obtained from the MATH3PL_EAP.SCO file that we created in our calibration. These $\hat{\theta}$ s, the response data, and the item parameter estimates are used to calculate the p_j s as well as the residuals ($x_{ij} - p_j(\hat{\theta}_i)$). Using the residuals, we calculate the 10 third-order partial correlations (i.e., Q_3^P). Table 6.4 shows the Q_3^P 's for the mathematics data example; the scatterplots

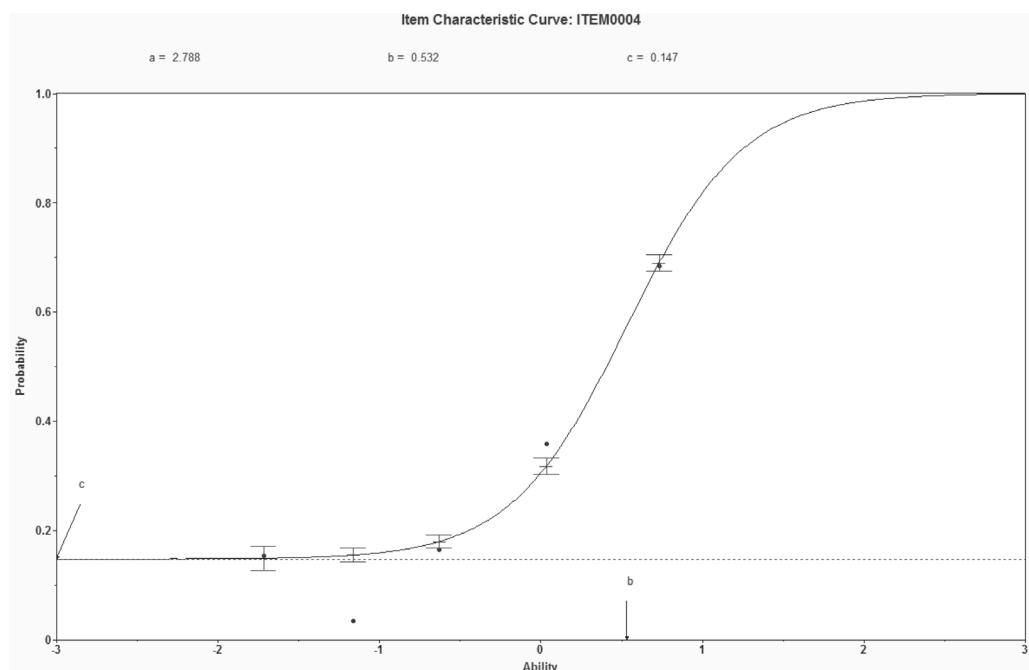
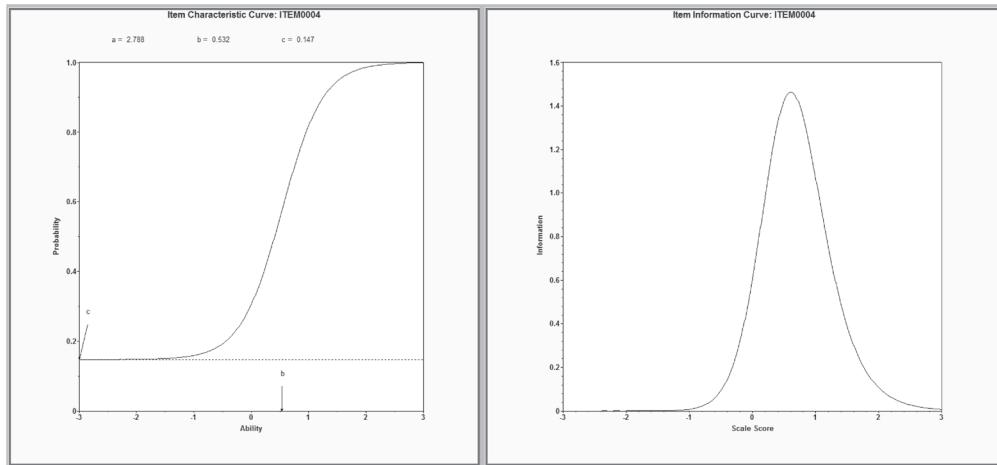


FIGURE 6.4. Empirical and predicted IRFs for item 4.

**FIGURE 6.5.** Item response and item information functions for item 4.

(not presented) corresponding to these values were inspected for anomalies, but none were found.

Figure 6.6 contains a dot density plot of our Q_3^P 's with the location of the mean \bar{Q}_3^P ($\bar{Q}_3^P = -0.29789$). To obtain \bar{Q}_3^P we use the Fisher r to \bar{z} transformation to convert each correlation to \bar{z} and then calculate the average \bar{z} . This average \bar{z} is transformed back to correlation. That is,

$$Q_3^P = \tanh\left(\sum_{j=1}^{L-1} \sum_{k=j+1}^L \text{arctanh}(Q_{3(j,k)}^P) / {}_L C_r\right), \quad (6.6)$$

where ${}_L C_r = (L(L-1))/2$.

To identify values that reflect item dependence, we use a “gap” approach informed by \bar{Q}_3^P in which an item pair that is substantially separated from the item pair cluster

TABLE 6.4. Q_3^P Statistics for the Math Data Set; $(Q_3^P)^2$ Are in Parentheses

		Items		
		1	2	3
2		-0.30300 (0.09181)		
3	4	-0.24700 (0.06101)	-0.44300 (0.19625)	
4	5	-0.23100 (0.05336)	-0.38500 (0.14823)	-0.37000 (0.13690)
5		-0.16900 (0.02856)	-0.28100 (0.07896)	-0.25100 (0.06300)
				-0.27800 (0.07728)

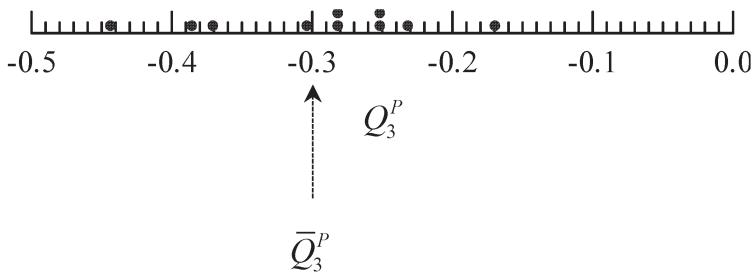


FIGURE 6.6. Dot density plot for Q_3^P .

reflects a potentially item dependent pair. For instance, we see that we have a cluster of values in the -0.30 to -0.22 range that contains \bar{Q}_3^P . These item pairs around the mean Q_3^P do not show item dependence. However, below this range there is a gap (~ 0.15) to get to the lowest value Q_3^P of -0.443 . This item pair (items 2 and 3) has almost 20% of their variability in common (Table 6.4). Conversely, above this range we have a gap (~ 0.07) to reach our rightmost point $Q_3^P = -0.169$ for item pair 1–5. However, this item pair has only about 3% of shared variability. As such, we do not consider this item pair to be exhibiting item dependence. Additionally, one might consider the two values, $Q_3^P = -0.37$ and $Q_3^P = -0.385$ for item pairs 3–4 and 2–4, respectively, with a gap of approximately 0.075 from the cluster's lowest value as potentially exhibiting dependence. These two pairs, items 3–4 and items 2–4, have 13.7% and 14.8%, respectively, of their variance in common. Although these three item pairs may be considered to be exhibiting item dependence, evidence of conditional dependence in the remaining seven pairs is absent. Our analysis shows that after fitting the unidimensional 3PL model to the data, the items in these three item pairs (i.e., items 2, 3, 4) had more than 13% of their residual variability in common.⁹ (These item pairs may or may not be found to be exhibiting item dependence with either the 1PL or 2PL models.)

How one deals with items that are considered sufficiently dependent to be problematic post administration is contingent on what one believes is the cause of the dependency. Again, inspection of the items exhibiting local dependence may be useful (the local dependence may be related to the locations of the items in the instrument, their text, dimensionality, and so on). In some cases where there is a great deal of dependence, it may be necessary to remove one of the dependent items for pragmatic reasons and because it is not clear as to why there is a dependency between the items. Because highly dependent items are in a sense redundant, the removal of one of the dependent items may not be problematic. In other cases, the items may be combined to form a testlet or the items combined to form an item parcel that is scored polytomously. In either case the instrument would need to be recalibrated.

For our example, the local dependence exhibited by the two item pairs could be addressed by forming a parcel for each pair. Parcel 1 would consist of items 1 and 4, whereas parcel 2 would involve items 2 and 5. Each parcel would have possible scores of 0 through 2, and the instrument would consist of three “items” (i.e., parcel 1, parcel 2,

and item 3). The corresponding response data could be calibrated using, for example, the polytomous partial credit model discussed in Chapter 7.

Fit Assessment: Model Comparison

In this and in previous chapters, our focus has been on whether a particular model is exhibiting model–data fit. We now present three model–data fit statistics that can be used for making model comparisons and selection. These complementary procedures should be used after obtaining evidence supporting model–data fit. The first of these is based on the likelihood ratio (G^2) test statistic for comparing the relative fit of hierarchically related models. The second statistic is analogous to the use of R^2 for comparing various regression models, whereas the third is based on an information criterion.

The change in G^2 across models can be used to determine whether two hierarchically related models significantly differ from one another. For instance, the 2PL model can be considered to be nested within the 3PL model because constraining the 3PL model's χ_j s to be 0 yields the 2PL model. Similarly, imposing the constraint that all items have the same discrimination parameter on the 2PL model produces the 1PL model. If we impose the constraints that all the items have a common item discrimination parameter and χ_j s equals 0, then the 3PL model reduces to the 1PL model. As such, the 1PL model is nested within both the 2PL and 3PL models. In the following discussion, we refer to the more complex (or less constrained) model as the *full* model and the less complex/simpler (or more constrained) model as the *reduced* model. The likelihood ratio test is the difference between two deviance statistics

$$\Delta G^2 = (-2 \ln L_R) - (-2 \ln L_F) = G_R^2 - G_F^2, \quad (6.7)$$

where L_R is the maximum of the likelihood for the reduced model and L_F is the maximum of the likelihood for the full model. The degrees of freedom (df) for evaluating the significance of ΔG^2 is the difference in the number of parameters between the full model and the reduced model.¹⁰ This statistic is distributed as a χ^2 when the sample size is large and the full (nesting) model holds for the data. A nonsignificant statistic indicates that the additional complexity of the nesting model is not necessary. For instance, if a comparison of the 2PL model with the 3PL model is not significant, then the additional estimation of the pseudo-guessing parameters (i.e., the increased model complexity) is not necessary to improve model–data fit over and above that obtained with the 2PL model.

Table 6.5 contains the values of the $-2 \log$ likelihoods (i.e., $-2\ln L$) for the 1PL, 2PL, and 3PL models from our BILOG calibrations. The $-2\ln L$ is the last entry from the converged solution's iteration history and is labeled `-2 LOG LIKELIHOOD:` in the output. As can be seen, as the models increase in complexity, the corresponding G^2 's decrease. The difference between the 1PL and 2PL models is

$$\Delta G^2 = (-2 \ln L_R) - (-2 \ln L_F) = 110,774.295 - 110,397.103 = 377.191$$

with 4 df . Therefore, at the instrument level the 2PL model represents a significant (at the 5% level) improvement in fit over the 1PL model. An analogous comparison between the 2PL and 3PL models also shows a significant improvement in fit by the 3PL model over the 2PL model. Therefore, the 3PL model fits significantly better than either the 2PL or 1PL model.

Our second model comparison statistic is complementary to ΔG^2 . This approach uses G^2 in a manner analogous to comparing various regression models' R^2 s. That is, the change in R^2 s may be used for assessing the relative improvement in the proportion of variability accounted for by one model over another model. In the current context, our strategy is to calculate the relative reduction in G^2 s (Haberman, 1978). For instance, for the comparison of the 2PL model (G_F^2) with the 1PL model (G_R^2) we would calculate

$$R_{\Delta}^2 = \frac{G_R^2 - G_F^2}{G_R^2} = \frac{110774.295 - 110397.103}{110774.295} = 0.0034$$

This R_{Δ}^2 indicates that the 2PL model results in a 0.34% improvement in fit over the 1PL model. Comparing the 3PL and 2PL models we have

$$R_{\Delta}^2 = \frac{110397.103 - 110066.023}{110397.103} = 0.0030$$

Therefore, using the more complex 3PL model results in an improvement of fit of 0.3% over the 2PL model. We do not consider this to be a meaningful improvement in fit vis-à-vis the increase in model complexity. Summarizing the results so far, we have that the 3PL model fits significantly better than the 2PL and 1PL models, but it does not result in a *meaningful* improvement of fit of over either model.

Table 6.5 shows the AIC and BIC values for the three models. As is the case with ΔG^2 above, we see that even taking the 3PL model's additional complexity into account (i.e., relative to the 1PL and 2PL models), these statistics indicate that it is the best fitting of these three models.

Although our triangulation with ΔG^2 , AIC, and BIC shows that the 3PL model is the best-fitting model of the three considered, our R_{Δ}^2 shows that the differences are slight. In fact, the correlation between the 1PL model-based $\hat{\theta}$ s and those of the 2PL model is 0.9908, and for the $\hat{\theta}$ s from the 2PL and 3PL models the correlation is 0.9869; the lowest

TABLE 6.5. Model Fit Statistics

Model	-2lnL	df	Relative Change	Number of Parameters	AIC	BIC
1PL ^a	110,774.295	25		6	110,786.295	110,833.595
2PL	110,397.103	21	0.0034	10	110,417.103	110,495.937
3PL	110,066.023	16	0.0030	15	110,096.023	110,214.273

^aFive item locations plus a common α .

correlation is between the 1PL and 3PL models' $\hat{\theta}$ s, $r = 0.9778$. That is, although the 3PL model is the best fitting of the three models, we have a high degree of linear agreement in the ordering of individuals across the three models. Based solely on the R^2_{Δ} , the magnitude of the $\hat{\theta}$ intercorrelations, the variability in the $\hat{\alpha}$ s (both this chapter and Chapter 5), and the axiom "Make everything as simple as possible, but not simpler" (Albert Einstein), we would select the 2PL model for modeling these data. (However, we believe that a reasonable argument can be made for selecting the 1PL model.) Additional points to consider in model selection are presented below in the section "Issues to Consider in Selecting among the 1PL, 2PL, and 3PL Models."

Example: Application of the 3PL Model to the Mathematics Data, MMLE, mirt

As in Chapter 5, we assume the data and the relevant libraries are loaded into our R workspace. To perform our calibration, we specify the 3PL model in our call to the `mirt` function (`ThreePL = mirt(mathdata, 1, '3PL', SE = T, SE.type = 'Fisher')`). Our calibration required 58 iterations to obtain convergence (Table 6.6).

Examining our item parameter estimates, we notice that our first item has a large $\hat{\chi}_1$ of 0.609 and a comparatively good discrimination ($\hat{\alpha}_1 = 2.289$); the item is located at $\hat{\delta}_1 = -0.703$. The large $\hat{\chi}_1$ coupled with good discrimination is somewhat counterintuitive. Our traditional item statistics corroborate the item's easiness, with almost 89% of the respondents providing a correct response (P-value = 0.8875); this item did not do as well as the other items in differentiating among observed scores (corrected $r_{1,NC} = 0.246$). Our observed score frequency distribution shows that only 3.5% of our sample have a $X = 0$, and our empirical IRFs for item 1 (Figure 6.7) show there is little observed data below approximately -1 . In toto, we conjecture that there is not enough information at the lower end of the continuum to "properly" estimate item 1's lower asymptote.

As Figure 6.7 shows, the leftmost empirical point ($\theta \cong -1.12$) has a proportion correct of about 0.33 that is not quite being captured by the IRF. With 6 fractiles we see a smoother empirical pattern that appears to indicate that the IRF should continue further down to reflect the leftmost empirical point. Because we are modeling the data, we decide to impose priors on our $\hat{\chi}$ s to enhance the correspondence between our $\hat{\chi}$ s and our data. This two-step process begins with determining the item parameter number for the parameter of interest and then specifying the prior using the item parameter number. To determine the item parameter number, we execute `mirt` using the `pars = 'values'` argument. The corresponding output object (`modThreePL`) is shown in Table 6.7, with the leftmost column containing the item parameter numbers and the column labeled `item` and `name` specifying the item and corresponding parameter label, respectively. For instance, item 1's information is on the first four lines, with item parameter number 1 being used for α_1 (labeled `a1`), number 2 for δ_1 (labeled `d1`), number 3 for χ_1 (labeled `g1`), and number 4 for γ_1 (labeled `u1`), respectively. We can identify the item parameter numbers for $\chi_1, \chi_2, \chi_3, \chi_4$, and χ_5 from the appropriate line in the `modThreePL` display.

TABLE 6.6. mirt Session for the 3PL Calibration of the Mathematics Data (No Prior)

```

> # read data, load mirt, etc.

> print((ThreePL = mirt(mathdata,1,'3PL',SE=T,SE.type='Fisher')))
Iteration: 58, Log-Lik: -55028.684, Max-Change: 0.00008

Calculating information matrix...

Call:
mirt(data = mathdata, model = 1, itemtype = "3PL", SE = T, SE.type = "Fisher")

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 58 EM iterations.
mirt version: 1.30
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Information matrix estimated with method: Fisher
Condition number of information matrix = 1897.529
Second-order test: model is a possible local maximum

Log-likelihood = -55028.68
Estimated parameters: 15
AIC = 110087.4; AICc = 110087.4
BIC = 110205.6; SABIC = 110157.9
G2 (16) = 55.71, p = 0
RMSEA = 0.011, CFI = NaN, TLI = NaN

> coef(ThreePL,simplify=TRUE,IRTpars=TRUE)
$items
      a      b      g      u
I1 2.289 -0.703 0.609 1
I2 2.640 -0.306 0.108 1
I3 2.523  0.154 0.212 1
I4 2.736  0.519 0.143 1
I5 1.618  0.867 0.154 1

> # get proportion correct
> summary(mathdata)
      I1          I2          I3          I4          I5
Min. :0.0000  Min. :0.0000  Min. :0.0000  Min. :0.0000  Min. :0.0000
1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
Median :1.0000 Median :1.0000 Median :1.0000 Median :0.0000 Median :0.0000
Mean   :0.8875 Mean   :0.644  Mean   :0.566  Mean   :0.427  Mean   :0.3873
3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:1.0000
Max.   :1.0000 Max.   :1.000  Max.   :1.000  Max.   :1.000  Max.   :1.0000

> # obtain corrected item total correlations & item total correlations
> NC=c(rep(-1,19601))                                     # create & initialize NC

> for (i in 1:19601){NC[i]=sum(mathdata[i,])}             # calculate NC

```

(continued)

TABLE 6.6. (*continued*)

```

> table(NC)                                     # frequency distribution
  NC
  0    1    2    3    4    5
  691 3099 4269 4116 4041 3385

> # calculate corrected point biserial & point biserial for each item
> for (j in 1:5){
+ print(j)
+ print(corr(NC-mathdata[,j]),mathdata[,j])); print(corr(NC,mathdata[,j])) }
[1] 1
[1] 0.2460252
[1] 0.4473804
[1] 2
[1] 0.4389591
[1] 0.6882788
[1] 3
[1] 0.4156754
[1] 0.6804917
[1] 4
[1] 0.4051335
[1] 0.6727627
[1] 5
[1] 0.3117375
[1] 0.6017258

> itemfit(ThreePL,S_X2$tables=T,empirical.table=1)   # item 1
$`theta = -1.1215`
      Observed  Expected z.Residual
cat_0     1309    554.1609   32.06538
cat_1      651  1405.8391  -20.13198

$`theta = -0.936`
      Observed  Expected z.Residual
cat_0     331    483.2472  -6.925715
cat_1    1629  1476.7528   3.961826
:

> itemfit(ThreePL,group.bins=10,empirical.plot=1,empirical.CI=0)  # produces Figure 6.7
> itemfit(ThreePL,group.bins=6,empirical.plot=1,empirical.CI=0)  # produces Figure 6.7

```

Alternatively, we can extract the parameter numbers by using a Boolean expression to select only the parameter numbers from the parnum variable when the variable name is set to g (with(modThreePL,parnum[name == 'g'])); we use the with function to minimize typing the output object name. The corresponding item parameter numbers for our $\chi_1, \chi_2, \chi_3, \chi_4$, and χ_5 are 3, 7, 11, 15, and 19, respectively.

To impose the prior, we specify the type ('prior.type') and its parameters ('prior_1', 'prior_2'). The prior has a location value of -1.5 and a scale value of 0.5. We use a generic for loop to implement the three assignments ('prior.type', 'prior_1', 'prior_2') for each item. Our for loop executes for each of our items; the number of items is stored in the output object ThreePL in the variable @Data\$nitems (see Table 6.6 for the creation of ThreePL). The body of the loop

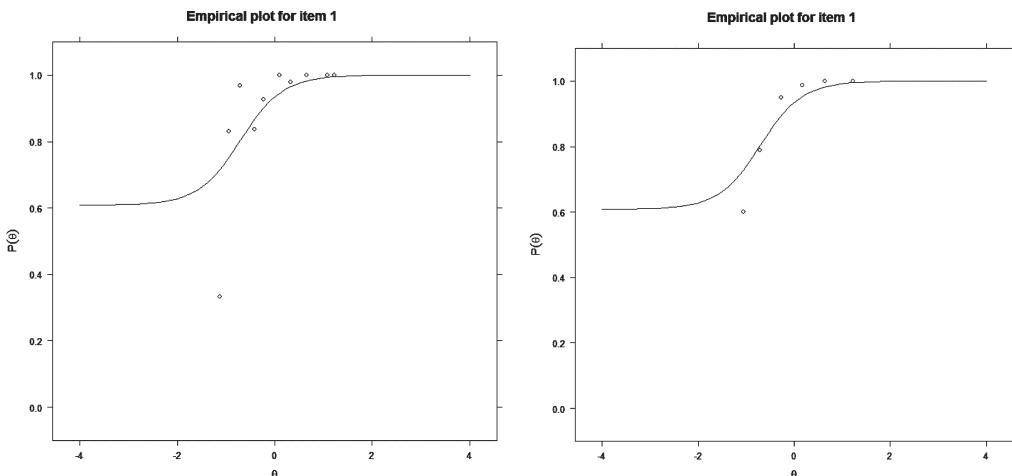


FIGURE 6.7. IRF for item 1 with observed proportions (left: 10 fractiles; right: 6 fractiles).

contains three assignments, with the indexing of the item parameter done by using the variable `itm`. We initialize `itm` to our first item parameter number 3. Each iteration of the `for` loop increments `itm` by the difference between successive parameter numbers (i.e., 4).

In our call to `mirt`, we pass our modified `modThreePL` by using the `pars` argument (`pars = modThreePL`). Our calibration required fewer iterations (i.e., 37) than when we did not impose priors on the estimation of the $\hat{\chi}$ s (i.e., 58 iterations). In contrast to our previous estimates for item 1, we now have $\hat{\alpha}_1 = 1.683$, $\hat{\delta}_1 = -1.542$, and $\hat{\chi}_1$ of 0.221. Our new $\hat{\chi}_1$ is closer to the data. Figure 6.8 shows the corresponding IRFs. Both figures show a better agreement with the data than is seen in Figure 6.7. Comparing items 2–5’s $\hat{\chi}$ s with and without use of the prior shows a difference on the order of 0.007 or less for items 3–5 and 0.058 for item 2. The items’ corresponding IRFs and item information are shown in Figure 6.9. The nonzero lower asymptotes are evident in the IRFs. Similarly, the varying discriminations is seen in both the slope of the IRFs and the heights of the item information functions. As in previous chapters, our `mirt` item parameter estimates show correlations of 0.991 or higher with those of BILOG.

Because our fit analysis at both the model and item levels proceeds as demonstrated in Chapter 5, we do not repeat it here. However, we note that according to the information criteria (e.g., AIC, BIC), the 3PL model is found to fit better than the 2PL model (i.e., `anova(TwoPL, ThreePL)`). Moreover, we show how to examine conditional dependence using Q_3 (`residuals(ThreePL, type = "Q3")`); Q_3 is one of the local dependency statistics provided by `mirt`. See Appendix G, “Conditional Independence Using Q_3 ,” for more information on using Q_3 .

We examine item parameter invariance using multiple-group analysis. We begin by creating two random samples, along with a binary group indicator variable (`grp`), followed by concatenating the two samples to create our data frame `mathdatagrp`. (Because we use Chapter 5’s seed (`set.seed(99999)`), these samples are the same as those in Chapter 5.) To impose priors on the $\hat{\chi}$ s, we use the `pars = 'values'`

TABLE 6.7. mirt Session for the 3PL Calibration of the Mathematics Data (Prior)

```

> # This is a continuation of the session from Table 6.6

> # go get item parameter numbers (3, 7, 11, 15, 19)
> print((modThreePL = mirt(mathdata,1,'3PL',SE=T,SE.type='Fisher',pars='values')))

  group item class name parnum      value lbound ubound    est prior.type prior_1
1   all   I1  dich  a1      1  0.8510000 -Inf     Inf  TRUE  none   NaN
2   all   I1  dich   d      2  2.3841111 -Inf     Inf  TRUE  none   NaN
3   all   I1  dich   g      3  0.1500000 0e+00      1  TRUE  none   NaN
4   all   I1  dich   u      4  1.0000000 0e+00      1 FALSE none   NaN
5   all   I2  dich  a1      5  0.8510000 -Inf     Inf  TRUE  none   NaN
6   all   I2  dich   d      6  0.7257898 -Inf     Inf  TRUE  none   NaN
7   all   I2  dich   g      7  0.1500000 0e+00      1  TRUE  none   NaN
:
18  all   I5  dich   d     18 -0.5626501 -Inf     Inf  TRUE  none   NaN
19  all   I5  dich   g     19  0.1500000 0e+00      1  TRUE  none   NaN
20  all   I5  dich   u     20  1.0000000 0e+00      1 FALSE none   NaN
21  all GROUP GroupPars MEAN_1      21  0.0000000 -Inf     Inf FALSE none   NaN
22  all GROUP GroupPars COV_11      22  1.0000000 1e-04    Inf FALSE none   NaN

prior_2
1   NaN
2   NaN
:
21  NaN
22  NaN

> with(modThreePL,parnum[name == 'g'])
[1] 3 7 11 15 19

> ThreePL@Data$nitems                                # ThreePL was created in Table 6.6
[1] 5

> item=3
> for(j in 1:ThreePL_a@Data$nitems[1]){
+   modThreePL[item,'prior.type']='norm'
+   modThreePL[item,'prior_1']=-1.5
+   modThreePL[item,'prior_2'] =0.5
+ item=item+4
+ } # end for j loop

> modThreePL                                         # checking that prior information was correctly imposed
  group item class name parnum      value lbound ubound    est prior.type prior_1
1   all   I1  dich  a1      1  0.8510000 -Inf     Inf  TRUE  none   NaN
2   all   I1  dich   d      2  2.3841111 -Inf     Inf  TRUE  none   NaN
3   all   I1  dich   g      3  0.1500000 0e+00      1  TRUE  norm  -1.5
4   all   I1  dich   u      4  1.0000000 0e+00      1 FALSE none   NaN
5   all   I2  dich  a1      5  0.8510000 -Inf     Inf  TRUE  none   NaN
6   all   I2  dich   d      6  0.7257898 -Inf     Inf  TRUE  none   NaN
7   all   I2  dich   g      7  0.1500000 0e+00      1  TRUE  norm  -1.5
:
19  all   I5  dich   g     19  0.1500000 0e+00      1  TRUE  norm  -1.5

:
> # use prior information with 'pars=' argument
> print((ThreePL = mirt(mathdata,1,'3PL',SE=T,SE.type='Fisher',pars=modThreePL)))
Iteration: 37, Log-Lik: -55033.198, Max-Change: 0.00008

```

(continued)

TABLE 6.7. (continued)

```

Calculating information matrix...
Warning message:
In ESTIMATION(data = data, model = model, group = rep("all", nrow(data)), :
  Information matrix with the Fisher method does not
  account for prior parameter distribution information

Call:
mirt(data = mathdata, model = 1, itemtype = "3PL", SE = T, SE.type = "Fisher",
      pars = modThreePL)

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 37 EM iterations.
mirt version: 1.30
M-step optimizer: nlmminb
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Information matrix estimated with method: Fisher
Condition number of information matrix = 50656.81
Second-order test: model is a possible local maximum

Log-posterior = -55033.2
Estimated parameters: 15
DIC = 110096.4
G2 (16) = 61.17, p = 0
RMSEA = 0.012, CFI = NaN, TLI = NaN

> print(coef(ThreePL, IRTpars=TRUE, printSE=T), digits=5)
$I1
      a      b      g      u
par 1.64259 -1.54157 0.22136  1
SE  0.11009  0.14091 0.02691 NA

$I2
      a      b      g      u
par 2.92857 -0.21561 0.16569  1
SE  0.27594  0.05758 0.05430 NA

$I3
      a      b      g      u
par 2.45749  0.14175 0.20510  1
SE  0.22644  0.04933 0.01449 NA

$I4
      a      b      g      u
par 2.88706  0.52929 0.15035  1
SE  0.32700  0.02762 0.04145 NA

$I5
      a      b      g      u
par 1.64094  0.87442 0.15803  1
SE  0.17635  0.04386 0.04594 NA

$GroupPars
  MEAN_1 COV_11
par      0      1
SE     NA     NA

```

(continued)

TABLE 6.7. (continued)

```

> anova(TwoPL,ThreePL) # The object TwoPL was created in Chapter 5

  Model 1: mirt(data = mathdata, model = 1, itemtype = "2PL", SE = T,
    SE.type = "Fisher")
  Model 2: mirt(data = mathdata, model = 1, itemtype = "3PL", SE = T,
    SE.type = "Fisher", pars = modThreePL)

      AIC      AICC     SABIC      HQ      BIC      DIC      logLik      logPost      df
  1 110417.0 110417.0 110464.0 110442.8 110495.8 110417.0 -55198.50 -55198.50  NaN
  2 110093.1 110093.1 110163.7 110131.8 110211.3 110096.1 -55031.55 -55033.03  5
  Bayes_Factor
  1          NA
  2          0

> itemfit(ThreePL,group.bins=10,empirical.plot=1,empirical.CI=0) # produces Figure 6.8 (left)
> itemfit(ThreePL,group.bins=6,empirical.plot=1,empirical.CI=0)  # produces Figure 6.8 (right)

> residuals(ThreePL,type="Q3") # Yen's Q3
  Q3 matrix:
    I1      I2      I3      I4      I5
  I1  1.000 -0.185 -0.100 -0.092 -0.059
  I2 -0.185  1.000 -0.300 -0.212 -0.137
  I3 -0.100 -0.300  1.000 -0.208 -0.097
  I4 -0.092 -0.212 -0.208  1.000 -0.158
  I5 -0.059 -0.137 -0.097 -0.158  1.000

> marginal_rxx(ThreePL)
[1] 0.6104937

> # empirical reliability
> ThreePLrxx=fscores(ThreePL,method="EAP",full.scores=T,full.scores.SE=T,returnER=T)
> ThreePLrxx
  F1
0.6294784

> # examination of invariance ----- use of priors
> set.seed(99999)
> caseU=runif(19601)
> sortmathdata=mathdata
> sortmathdata$unif=caseU
> sortmathdata=sortmathdata[order(sortmathdata$unif),]
> mathdata1=sortmathdata[1:9800,] ; mathdata2=sortmathdata[9801:19601,]
> mathdata1=within(mathdata1,rm(unif)) ; mathdata2=within(mathdata2,rm(unif))
> names(mathdata1) = c(paste0("I",1:5)) ; names(mathdata2) = c(paste0("I",1:5))

> mathdata1$grp='0' ; mathdata2$grp='1' # create {0,1} group variable

> mathdatagrp=rbind(mathdata1, mathdata2) # concatenate the two randomized groups
> grpvar=mathdatagrp$grp # extract group variable

> mathdatagrp=within(mathdatagrp,rm(grp))
> ThreePLgrp=multipleGroup(mathdatagrp,1,itemtype='3PL',group=grpvar, pars='values')
> ThreePLgrp

> ThreePLgrp
  group item  class name parnum      value lbound ubound    est prior.type prior_1
  1     0   I1 dich   a1      1  0.8510000  -Inf     Inf  TRUE    none    NaN
  2     0   I1 dich     d      2  2.3841111  -Inf     Inf  TRUE    none    NaN
  3     0   I1 dich     g      3  0.1500000  0e+00      1  TRUE    none    NaN
  4     0   I1 dich     u      4  1.0000000  0e+00      1 FALSE   none    NaN

```

(continued)

TABLE 6.7. (continued)

```

5      0   I2    dich    a1      5  0.8510000  -Inf     Inf  TRUE      none   NaN
6      0   I2    dich     d      6  0.7257898  -Inf     Inf  TRUE      none   NaN
:
41     NaN
42     NaN
43     NaN
44     NaN

> itm=with(ThreePLgrp,parnum[name == 'g'])                                #obtain & store item numbers for 'g'
> itm
[1] 3 7 11 15 19 25 29 33 37 41
> ngrps=2                                                               # number of groups

> # imposing priors on group "0" and group "1"
> for(j in 1:(ThreePL@Data$nitems[1]*ngrps)){
> + ThreePLgrp[item[j], 'prior.type']='norm'; ThreePLgrp[item[j], 'prior_1']=-1.5;
>     ThreePLgrp[item[j], 'prior_2']=0.5      }

> ThreePLgrp
# checking that prior information was correctly imposed
  group item  class name parnum      value lbound ubound    est prior.type prior_1
1      0   I1    dich  a1      1  0.8510000  -Inf     Inf  TRUE      none   NaN
2      0   I1    dich    d      2  2.3841111  -Inf     Inf  TRUE      none   NaN
3      0   I1    dich    g      3  0.1500000  0e+00     1  TRUE    norm  -1.5
4      0   I1    dich    u      4  1.0000000  0e+00     1 FALSE    none   NaN
5      0   I2    dich    a1      5  0.8510000  -Inf     Inf  TRUE      none   NaN
6      0   I2    dich    d      6  0.7257898  -Inf     Inf  TRUE      none   NaN
7      0   I2    dich    g      7  0.1500000  0e+00     1  TRUE    norm  -1.5
8      0   I2    dich    u      8  1.0000000  0e+00     1 FALSE    none   NaN
:
41     0.5
42     NaN
43     NaN
44     NaN

> ThreePLgrp=multipleGroup(mathdatagrp,1,itemtype='3PL',group=grpvar,pars=ThreePLgrp)
Iteration: 40, Log-Lik: -55030.224, Max-Change: 0.00010

> ThreePLgrp
Call:
multipleGroup(data = mathdatagrp, model = 1, group = grpvar,
itemtype = "3PL", pars = ThreePLgrp)

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 40 EM iterations.
mirt version: 1.30
M-step optimizer: nlminb
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Log-posterior = -55030.22
Estimated parameters: 30
DIC = 110120.4
G2 (1) = 79.67, p = 0
RMSEA = 0.063, CFI = NaN, TLI = NaN
← Nparm*L^2 groups

```

(continued)

TABLE 6.7. (continued)

```

> coef(ThreePLgrp, simplify=TRUE, IRTpars=TRUE)
$`0`                                         ← Group 0 item parameter estimates
$items
    a      b      g u
I1 1.629 -1.584 0.198 1
I2 2.977 -0.208 0.185 1
I3 2.474  0.130 0.212 1
I4 3.184  0.527 0.158 1
I5 1.661  0.875 0.162 1

$means
F1
0

$cov
F1
F1  1

$`1`          Group 1 item parameter estimates
$items
    a      b      g u
I1 1.637 -1.565 0.198 1
I2 2.931 -0.210 0.155 1
I3 2.429  0.151 0.197 1
I4 2.644  0.532 0.143 1
I5 1.622  0.875 0.155 1

$means
F1
0

$cov
F1
F1  1

> plot(ThreePLgrp, type = 'trace', theta_lim=c(-4,4))           # Figure 6.10
> plot(ThreePLgrp, type="score",theta_lim=c(-4,4))            # Figure 6.11
>
> # end of examination of invariance -------

> plot(ThreePL, type = 'trace', theta_lim=c(-4,4))             # produces Figure 6.9 (top)
> plot(ThreePL, type = 'infotrace', theta_lim=c(-4,4))        # produces Figure 6.9c (top)

```

argument with the `multipleGroup` function. However, we store the item parameter numbers for the two groups in `itm` to directly access them in our for loop (`itm = with(ThreePLgrp, parnum[name == 'g'])`). Group 0's item parameter numbers are the same as above, whereas for group 1 we have 25, 29, 33, 37, and 41. We impose the priors on our two groups using a for loop in which the appropriate item numbers are indexed by `j` (i.e., `itm[j]`). We call `multipleGroup` a second time and pass to it our data frame containing the prior information (`pars = ThreePLgrp`). Convergence was achieved in 40 iterations. Comparison of group 0's item parameter estimates with those of group 1's shows close correspondence. Our groups' IRFs for each item are shown in Figure 6.10. At the item level, the agreement between the two sets of IRFs for

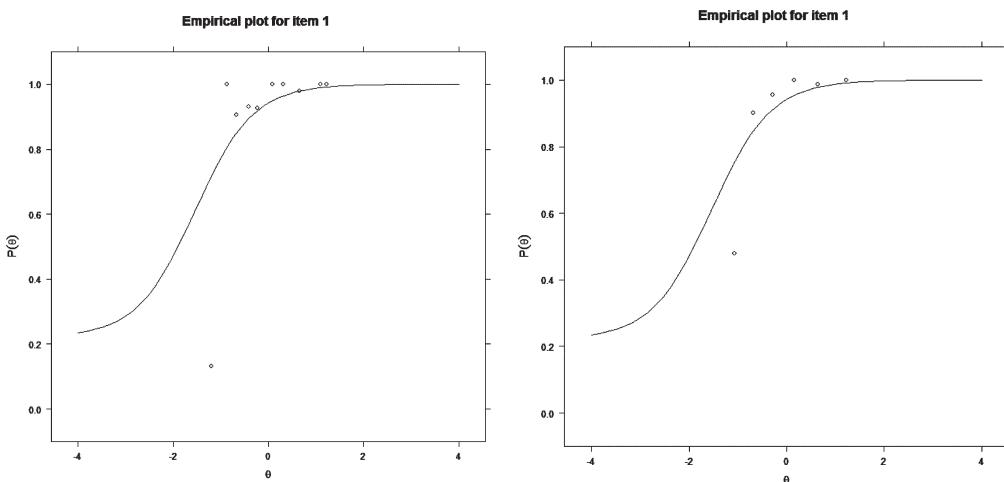


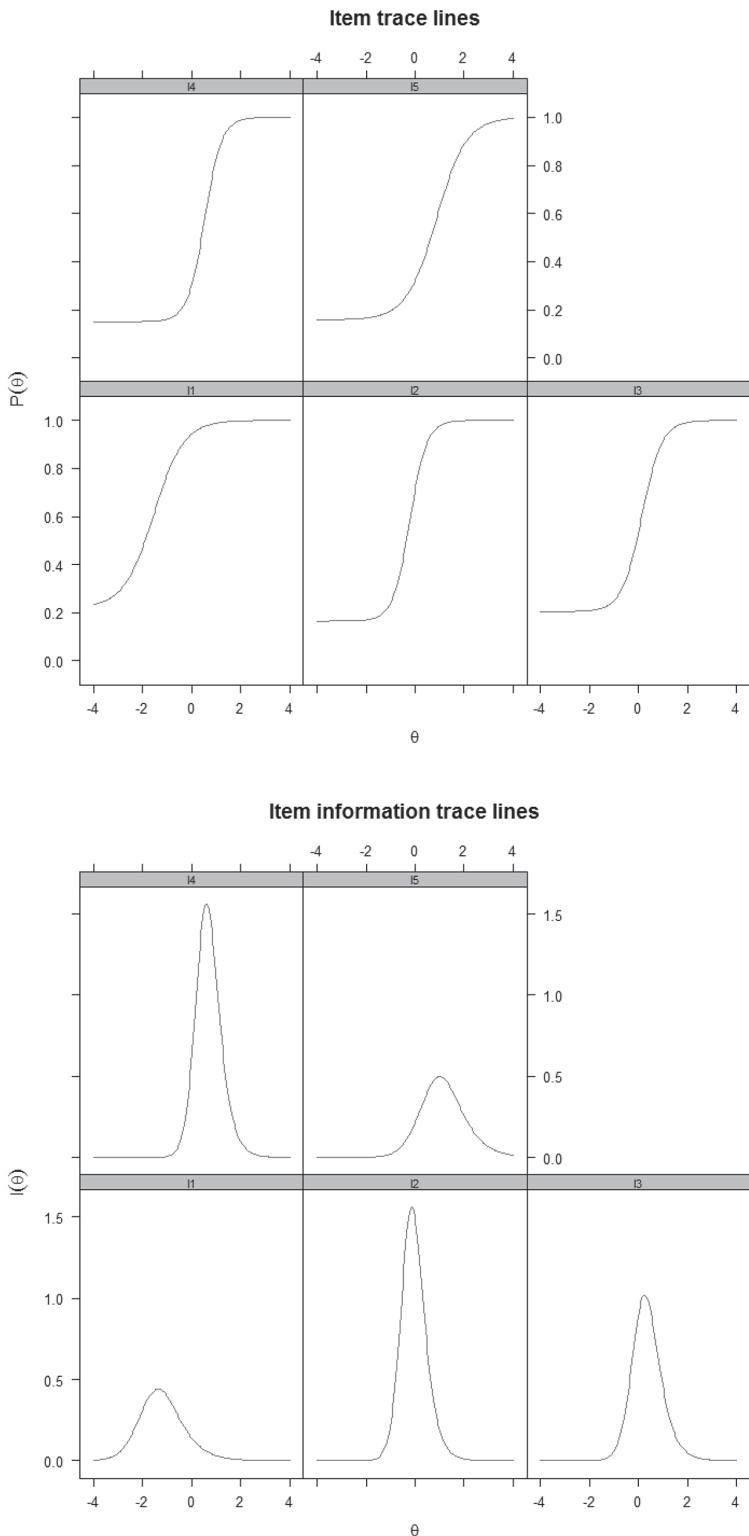
FIGURE 6.8. IRF for item 1 with observed proportions (left: 10 fractiles; right: 6 fractiles).

each item provides us with evidence of invariance. (Because our item parameter point estimates show estimation error, we interpret the minor discrepancies between IRFs as being within this margin of error.) Additionally, at the “model level,” our groups’ total characteristic curves (TCCs) show strong agreement with one another and provide us with additional invariance evidence (Figure 6.11).

Assessing Person Fit: Appropriateness Measurement

Various person fit measures have been previously discussed. From one perspective, these measures are trying to determine whether the person is behaving in a fashion consistent with the model. Alternatively, one may ask, what is the *appropriateness* of a person’s estimated location, $\hat{\theta}$, as a measure of their true location (θ)? For instance, imagine that a person has a response pattern of missing easy items and correctly answering more difficult items. Did this pattern arise from the person’s correctly guessing on some difficult items and incorrectly responding to easier items, or does this reflect a person who was able to copy the answers on some items? Various statistically based indices have been developed to measure the degree to which an individual’s response pattern is unusual or is inconsistent with the model used for characterizing their performance. These indices of person fit are examples of *appropriateness measurement* (e.g., Levine & Drasgow, 1983; Meijer & Sijtsma, 2001).

One index, l_z , has been found to perform better than other person fit measures (e.g., Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, & Williams, 1985). This index is based on a standardization of the person log likelihood function to address the interaction of $\ln L$ and θ . As such, this standardization of log likelihood allows us to compare individuals at different θ levels on the basis of their l_z values.

**FIGURE 6.9.** IRFs and item information for all items.

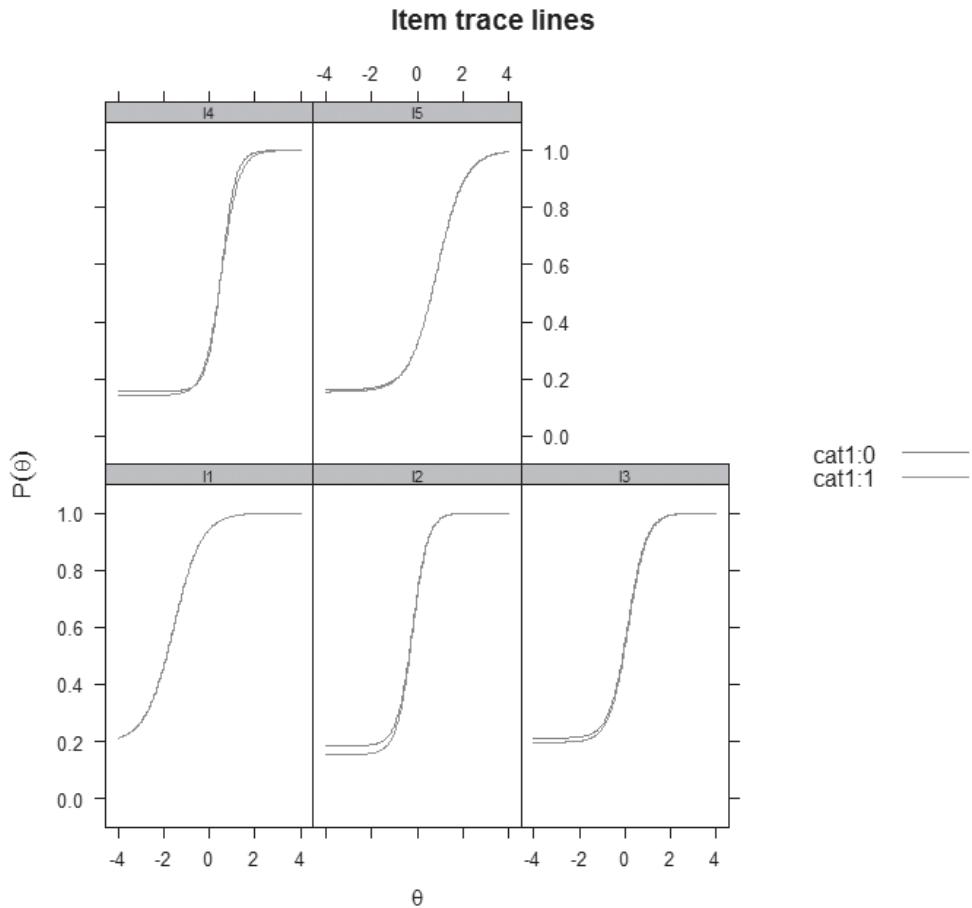


FIGURE 6.10. IRFs for two-group analysis.

To present l_z we start with the log likelihood function for a person i 's response vector

$$\ln L(\underline{x}_i, \theta, \underline{\alpha}, \underline{\delta}, \underline{\chi}) = \sum_{j=1}^L [x_{ij} \ln(p_j) + (1-x_{ij}) \ln(1-p_j)] \quad (6.8)$$

To standardize $\ln L$ we need both its variance and expected value. The expected value of the $\ln L$ is given by

$$\mathcal{E}(\ln L) = \sum_{j=1}^L [p_j \ln(p_j) + (1-p_j) \ln(1-p_j)] \quad (6.9)$$

and its variance by

$$Var(\ln L) = \sum_{j=1}^L \left\{ p_j (1-p_j) \left[\ln \frac{p_j}{1-p_j} \right]^2 \right\} \quad (6.10)$$

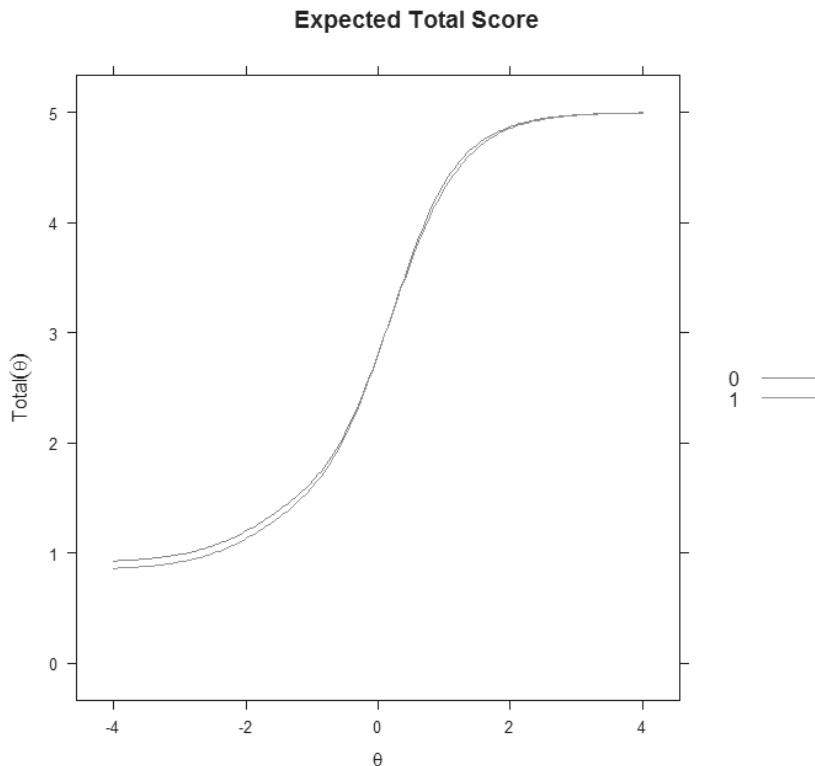


FIGURE 6.11. Total characteristic curves for two-group analysis.

Using Equations 6.8–6.10 and the z -score formula, we obtain

$$l_z = \frac{\ln L - \mathcal{E}(\ln L)}{\sqrt{\text{Var}(\ln L)}} \quad (6.11)$$

In practice, we use estimates in lieu of parameters in the calculation of p_j (e.g., $\hat{\theta}$ for θ).

Although l_z is purported to have a unit normal distribution, this has not necessarily been true for instruments of different lengths (Drasgow et al., 1985, 1987; Levine & Drasgow, 1983). Moreover, because the l_z uses person parameter estimates in its calculation, it is not asymptotic normal. To this end, Nering (1995) found l_z 's detection accuracy approaching the significance level is adversely affected by how well the person locations are estimated. Therefore, using the standard normal curve for hypothesis testing with l_z may be inadvisable in some situations. Nevertheless, various guidelines exist for using l_z for informed judgment. In general, a “good” l_z is one around 0.0. A negative l_z reflects a relatively unlikely response vector (i.e., inconsistent responses), whereas a positive value indicates a comparatively more likely response vector than would be expected on the basis of the model (i.e., hyperconsistent responses). Also see

Appendix G, “The Person Response Function,” for a graphical approach that can be used for detecting aberrant response vectors.

Snijders (2001) proposed an alternative to l_z that addresses the use of person estimates in its calculation. Thus, Snijders modifies l_z by incorporating a set of modification weights (\tilde{w}_j) in calculating l_z^*

$$l_z^* = \frac{\ln L - \mathcal{E}(\ln L) + c_L r_0}{\sqrt{Var^*(\ln L)}}, \quad (6.12)$$

where $Var^*(\ln L) = \sum_{j=1}^L \{p_j(1-p_j)w_j^{*2}\}$, $w_j^* = w_j - c_n r_j$, $c_L = \sum_{j=1}^L p'_j w_j / \sum_{j=1}^L p'_j r_j$

(p'_j is the first derivative of the model), $w_j = \ln[p_j/1-p_j]$, r_j depends on the model, and r_0 depends on the ability estimation technique and model. For example, for the 1PL model $r_j = 1$, for the 2PL model $r_j = \alpha_j$, and the 3PL $r_j = [\alpha_j \exp(\theta - \delta_j)]/[\chi_j + \alpha_j \exp(\theta - \delta_j)]$ with $r_0 = 0$ for MLE and $r_0 = -\theta$ for MAP and a θ distribution that is $N(0,1)$ (Magis, Raîche, & Béland, 2012).

The R package `PerFit` (Tendeiro, Meijer, & Niessen, 2016, 2018) can be used to calculate l_z and l_z^* as well as other person fit statistics. Table 6.8 shows our R session for obtaining l_z^* for our math data calibration using `mirt`. We begin by extracting our item parameter estimates into the object `itests` and then convert our person estimates to a vector `PersonEst`. Both are passed to the `lzstar` function to calculate each person's l_z^* with the results stored in the object `lzstar_stat`. We then use the `cutoff` function to obtain a screening value for the 5% level (`Blvl = .05`); `cutoff` uses 1000 bootstraps to generate an empirical sampling distribution and determine the value that cuts off 5% of the distribution. Passing the l_z^* results (`lzstar_stat`), and the `cutoff` function's output object (`lzstarcut_05`) to the `flagged.resp` function allows us to create an object that contains those cases whose $|l_z^*|$ values exceed the absolute value of screening point (`$Cutoff = -1.8083`). For our example, 562 cases (2.87% of our sample; `Prop.flagged`) are identified as having $|l_z^*|$ values exceeding the $|\text{screening point}|$ value.

Figure 6.12 contains the distribution of l_z^* values, with the vertical line indicating the screening value's location (-1.808) along with its confidence band (CB) on the abscissa. Thus, cases to the left of the vertical line are potentially misfitting persons. Alternatively, the lower bound of the CB could be used if one wishes to take into estimation error in identifying individuals for further examination. In this latter case, the tick marks on the top of the graph identify these potentially misfitting persons. Below we examine some cases from this distribution.

From above we know that items 1–5's P-values are 0.887, 0.644, 0.566, 0.427, and 0.387, respectively. Thus, for the first case identified (`FlaggedID = 20`—the 20th line in the data file), they incorrectly answered the easiest and hardest items as well as an item of moderate difficulty (`x = 01010`). In contrast, the case with the `FlaggedID` of 19306 incorrectly answered the easiest item, but correctly answered the progressively more

TABLE 6.8. PerFit Session to Obtain l_z^* for the 3PL Calibration of the Mathematics Data (Prior)^a

```

> library(PerFit)
> packageVersion("PerFit")
[1] '1.4.3'

> itests=coef(ThreePL,simplify=TRUE,IRTpars=TRUE)$items
  [,c('a','b','g')] # extract estimates
> PersonEst=as.vector(peopleThreePL[,1])
> lzstar_stat=lzstar(mathdata, IRT.PModel = "3PL",Ability=PersonEst,IP=itests)

> # determine screening value for the 5% level
> lzstarcut_05=cutoff(lzstar_stat,ModelFit="Parametric", Blvl=.05)
> FlgdCase_lzstar = flagged.resp(lzstar_stat,cutoff.obj=lzstarcut_05,scores=T)
> FlgdCases=FlgdCase_lzstar$Scores
> FlgdCase_lzstar$Cutoff
$Cutoff
[1] -1.8083

$Cutoff.SE
[1] 0.1751

$Prop.flagged
[1] 0.0287

$Tail
[1] "lower"

$Cutoff.CI
  2.5%   97.5%
-1.9985 -1.4599

attr(,"class")
[1] "PerFit.cutoff"

> FlgdCases=FlgdCase_lzstar$Scores
> head(FlgdCases,6)
  FlaggedID It1 It2 It3 It4 It5 PFscores
[1,]      20   0   1   0   1   0  -2.4340
[2,]      35   0   1   1   0   0  -1.7903
[3,]      41   1   0   1   1   1  -1.6627
[4,]      45   0   1   1   1   0  -2.9155
[5,]      55   0   1   1   1   0  -2.9155
[6,]     114   1   0   1   1   1  -1.6627

> tail(FlgdCases,4)
  FlaggedID It1 It2 It3 It4 It5 PFscores
[559,]    19306   0   1   1   1   1  -4.0532
[560,]    19354   0   1   1   0   1  -2.7950
[561,]    19525   0   1   1   0   1  -2.7950
[562,]    19533   0   1   1   1   0  -2.9155

> # for consistency with the above results we pass the cutoff object to the
> # plot function otherwise it will recompute the screening value
> plot(lzstar_stat,cutoff.obj=lzstarcut_05,Type="Histogram") # produces Figure 6.12

> PRFplot(mathdata,respID=19601,IP=itests,Ability=PersonEst) # produces Figure 6.13
  Respondent 19601: Press ENTER.

> PRFplot(mathdata,respID=19306,IP=itests,Ability=PersonEst) # produces Figure 6.13
  Respondent 19306: Press ENTER.

```

^aThe cutoff function uses a bootstrap to determine the screening value. The bootstrap uses a random number generator. The seed used is 88888.

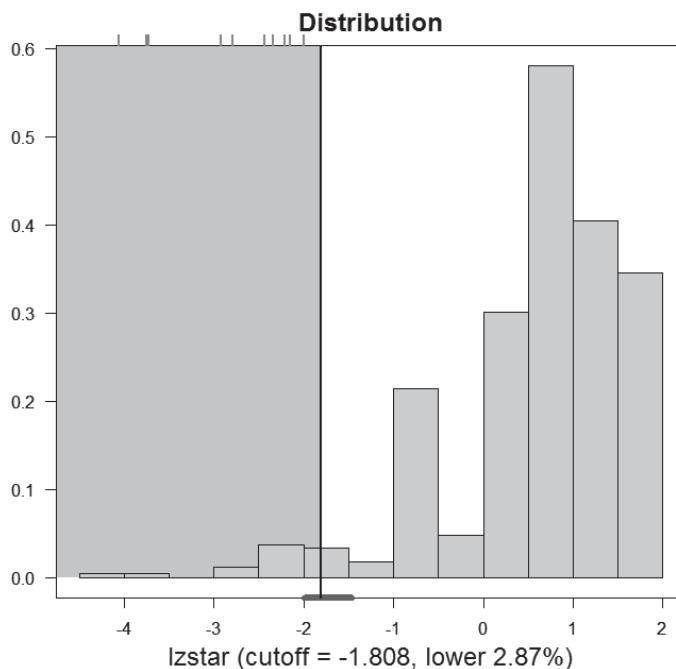


FIGURE 6.12. Distribution of person fit scores.

difficult items. Both of these individuals are not behaving consistently with expectations. As such, their $\hat{\theta}$ s may be inappropriate for them. Of course, with only five items we have a small item sample and insufficient information to fully determine if these are a problem. For example, if we had 20 items, we might very well find that 19,306 would behave consistent with expectations.

With five items we can easily look at the response pattern. However, with a longer instrument a graphical approach may be more instructive. Therefore, to demonstrate a graphical approach, we use the PRFplot function to obtain the person response function (PRF). As discussed in Appendix G, the PRF relates the probability of a response of 1 to item location. In Figure 6.13, we show the PRFs for two persons. The left side is for a person (#19,601) who was not identified as potentially misfitting. This person's response pattern of 11110 (i.e., correctly answering all items except the most difficult) and PRF are consistent with what is expected. That is, as items become more difficult relative to this person's $\hat{\theta}$, the probability of a correct response decreases; $\hat{\theta}_{19601} = 0.720$. In contrast, person #19,306 ($\mathbf{x} = 01111$; $\hat{\theta}_{19306} = 0.404$) has a PRF that is inconsistent with what one would expect using the 3PL model. One possible explanation of this \mathbf{x} is that this person's initial inattentiveness/carelessness may have led to their incorrect response to the first item, although they had the ability to correctly answer the easiest item. Thus, this person should have had a $\mathbf{x} = 11111$ with a commensurate $\hat{\theta}$ of 1.233. Alternatively, it may be that the person guessed and/or copied some or all of their responses to items 2–5. In this case, their \mathbf{x} should be something along the lines of 00000, with an $\hat{\theta} = -1.396$. It is also possible that there is something unique about

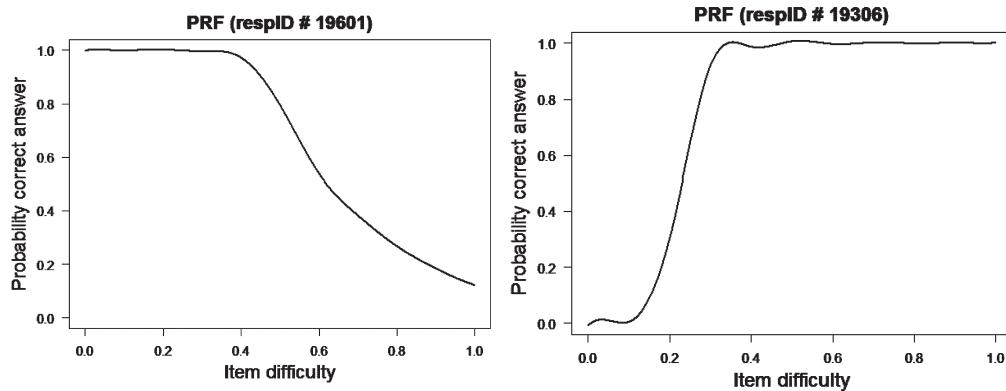


FIGURE 6.13. Person fit plots for fitting person (left) and misfitting person (right).

item 1 that led this person to respond incorrectly. In other words, the $\hat{\theta}_{19306}$ of 0.404 is an inappropriate estimate of this person's math ability. Of course, we have insufficient information to determine the cause of this person's response pattern.

Information for the Three-Parameter Model

The amount of information an item provides for estimating θ under the 3PL model is

$$I_j(\theta) = \alpha_j^2 \left[\frac{(p_j - \chi_j)^2}{(1 - \chi_j)^2} \right] \left[\frac{1 - p_j}{p_j} \right]. \quad (6.13)$$

Because guessing behavior reflects "noise," it may be intuited that one effect of a nonzero χ_j is to reduce the amount of information available for locating people on the θ continuum.^{10,11} Equation 6.13 shows that this is indeed the case. For a given α_j and δ_j , an item provides more information for person estimation when $\chi_j = 0$ than when it is nonzero. Therefore, for the 3PL model the upper limit of $I_j(\theta)$ is given by the more restrictive 2PL model. If one sets $\chi_j = 0$ and simplifies, then Equation 6.13 reduces to Equation 5.4.

In contrast to the 1PL and 2PL models with their maximum item information at δ_j , Figure 6.5 shows that for the 3PL model the peak of the item information does not occur at δ_j but slightly above it. This offset from δ_j is given by¹²

$$\frac{\ln \left[\frac{1}{2} + \frac{\sqrt{1+8\chi_j}}{2} \right]}{\alpha_j}.$$

At this location, the maximum item information value is (Lord, 1980)

$$\text{Max}(I_j(\theta)) = \frac{\alpha_j^2}{8(1-\chi_j)^2} [1 - 20\chi_j - 8\chi_j^2 + (1+8\chi_j)^{1.5}]. \quad (6.14)$$

As has previously been the case, the total information for an instrument is the sum of the item information

$$I(\theta) = \frac{1}{\sigma_e^2(\theta)} = \sum_{j=1}^L I_j(\theta). \quad (6.15)$$

In the foregoing, we have focused on the amount of information an item provides for estimating a person's location.^{13,14} However, we can also look at how much information the calibration sample provides for estimating a particular item parameter. The information for estimating α_j , δ_j , and χ_j is, respectively (Lord, 1980).

$$I_{\alpha_j} = \frac{1}{(1-\chi_j)^2} \sum_i^N \left((\theta_i - \delta_j)^2 (p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.16)$$

$$I_{\delta_j} = \frac{\alpha_j^2}{(1-\chi_j)^2} \sum_i^N \left((p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.17)$$

and

$$I_{\chi_j} = \frac{1}{(1-\chi_j)^2} \sum_i^N \frac{(1-p_j)}{p_j}. \quad (6.18)$$

In Figure 6.14, we present the information function (dash line) for estimating α_5 with item 5's IRF (solid line) overlaid.¹⁵ As can be seen, the information function is bimodal with different maxima. These different maxima reflect that we have a nonzero χ_5 . As χ_j increases, the left maximum decreases and shifts its location, whereas the right maximum increases in value and stays at the same θ location. The modes are located in the θ neighborhood of the IRF beginning its trajectory toward becoming asymptotic. It is also apparent that the modes occur on opposite sides of the item's location, with the leftmost mode always less than the rightmost mode. This characteristic is a reflection of positive α_5 (i.e., if $\alpha_5 < 0$, then the leftmost mode would be greater than the rightmost mode). As α_j decreases, the distance between the modes increases, the maxima values increase, and function broadens. The location of the minimum (i.e., 0) of the information function between the two modes corresponds to δ_5 . In other words, persons located at the item's location do not contribute information for estimating the item's discrimination.

Figure 6.15 shows that the information function for estimating δ_j is unimodal, with the mode located at the item's δ_j . Therefore, individuals around the item's location provide the greatest information for estimating δ . As is the case with Figure 6.14, the different heights of the modes across the items is a reflection of the interaction among the item's parameters as well as their different values across items. In short, poor estimation of one or more of the parameters (e.g., χ_j) affects the estimation of the item's other parameter(s).

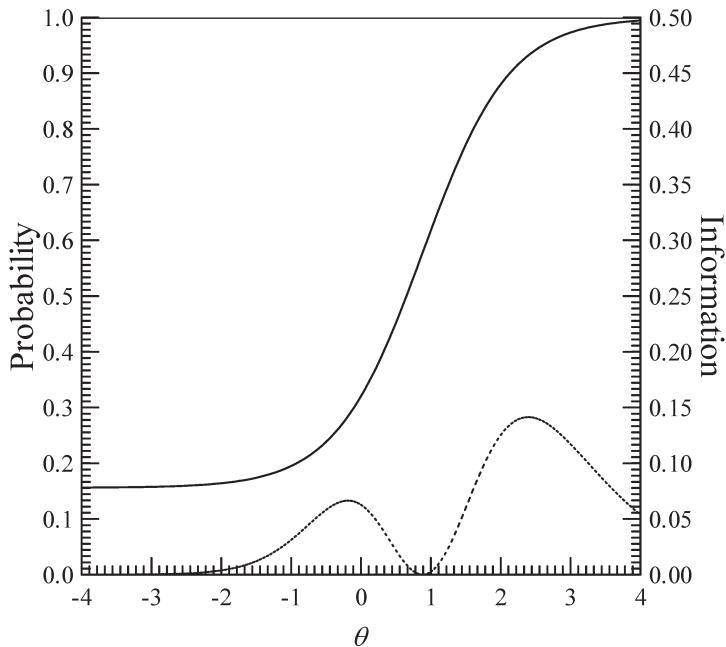


FIGURE 6.14. Information for estimating α_j as a function of θ for item 5 ($\hat{\alpha}_5 = 1.608$, $\hat{\delta}_5 = 0.883$, $\hat{\chi}_5 = 0.156$).

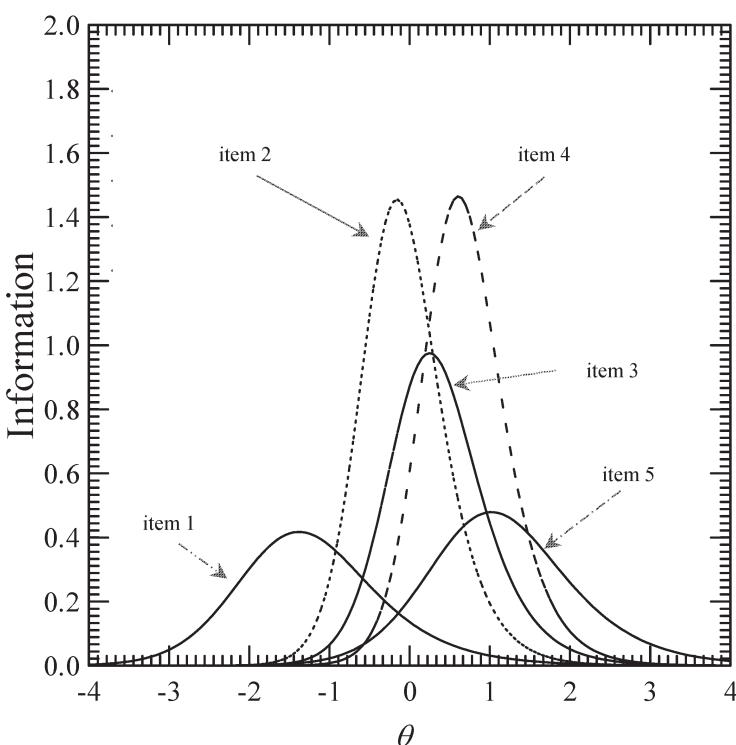


FIGURE 6.15. Information for estimating δ_j as a function of θ for each of five items.

With respect to χ_j one sees (Figure 6.16) that most information for estimating χ_j comes from the lower end of the θ continuum. Depending on the particular item, there is virtually no useful information for estimating χ_j from individuals located above 2.0. However, the information functions' plateaus show that even at the lower end of the θ continuum there is a finite amount of information available for estimating χ_j . Moreover, the larger the χ_j , the greater the shift in the beginning of this plateau toward the lower end of the continuum than when χ_j is smaller. We also see that the larger the χ_j , the lower the plateau, indicating less information for estimating these large χ_j values than for estimating smaller χ_j values.

For completeness, we now discuss the information functions for the 1PL and 2PL models. If we plot the information functions for estimating δ_j for the 1PL model, we find that across items the corresponding information functions have a constant height, with the location of the modes corresponding to the items' δ_j s. In addition, the information functions for estimating a common α across items are bimodal, but unlike the 3PL model case, the functions have a constant height across modes and across items. The minima of the information functions are zero and occur between the two modes at the items' δ_j s.

For the 2PL model, the information function for estimating the δ_j is also unimodal. Its height across items varies as a direct function of the items' α_j s, with the location of the modes corresponding to an item's δ_j . With respect to item discrimination, the information function for estimating α_j is bimodal, with a constant height across the modes for an item and equidistant from δ_j . However, the modes vary across the items as an

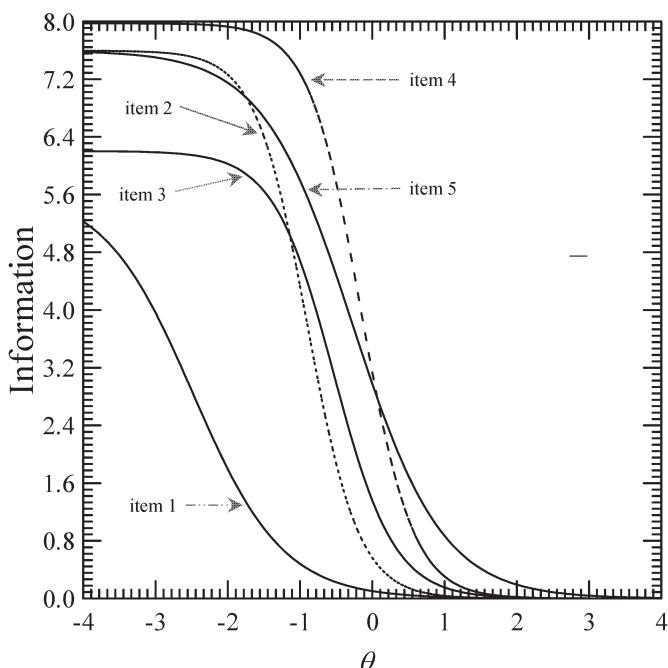


FIGURE 6.16. Information for estimating χ_j as a function of θ for each of five items.

indirect function of the items' α_j s. As is the case with the 1PL and 3PL models, the location of the minimum of the information function between the two modes corresponds to the item's δ_j and has a value of 0.

Metric Transformation, 3PL Model

Linear rescaling of α_j and δ_j (or their estimates) is accomplished as performed with the 2PL model. Because the pseudo-guessing parameter is on the probability scale, it does not have an indeterminacy in its scale and there is no need to rescale χ_j . Person location parameters (or their estimates) are transformed by $\theta^* = \zeta(\theta) + \kappa$.

The total characteristic curve for the 3PL model is determined as shown, for example, in Chapter 4. As is the case with the 1PL and 2PL models, all individuals with the same location, θ , obtain the same expected trait score, T . Furthermore, neither θ nor T depends on the distribution of persons. However, unlike the case with the previous models, with the 3PL the TCC lower asymptote is asymptotic with $\sum \chi_j$. As an example, the expected trait score for individuals with a $\hat{\theta}$ of 1.1746 on our mathematics test would be

$$T = \sum_{j=1}^L p_j = 0.9927 + \dots + 0.6813 = 4.4952.$$

Therefore, a person with an estimated location of 1.1746 would be expected to obtain almost 4.5 correct answers on the mathematics test. Figure 6.17 contains the TCC with the transformation of $\hat{\theta} = 1.1746$ to its corresponding T identified. Comparing this figure with the TCC for the 1PL model (Chapter 4, Figure 4.9) shows that it is steeper than the 1PL model's. The steepness of the TCC is a function of not only the discrimination parameter estimates (for the 3PL model the mean α is 2.3778 and for the 1PL model the common α is 1.421), but also the variability of the δ_j s as well as the magnitude of the χ_j s. As is seen, the lower asymptote of the TCC approaches the $\sum \chi_j = 0.889$, and its upper asymptote is the instrument's length because $\gamma_j = 1$ for all IRFs.¹⁶

Handling Missing Responses

From the preceding discussion, we know that IRT models are concerned with modeling *observed* responses. However, in working with empirical data, one will, at times, encounter situations where some items do not have responses from all individuals in the calibration sample. Some of these missing data may be considered to be missing by design or may be structurally missing. For example, one may administer an instrument to one group of people and an alternate form of the instrument to another group. If these two forms have some items in common, then the calibration sample can consist of both groups. As a result, our data contain individuals who have not responded to all items. Figure 11.1 in Chapter 11 contains a graphical depiction of this. In situations where the nonresponses are missing by design, these missing data may be ignored because of the IRT properties of person and item parameter invariance. However, when non-

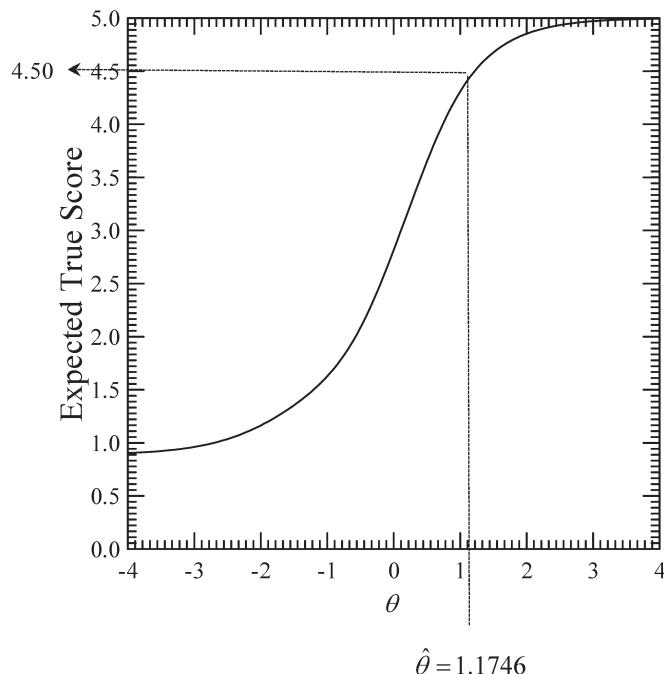


FIGURE 6.17. TCC for the five-item mathematics instrument calibrated with the 3PL model.

responses are not structurally missing, then one needs to consider how to treat these nonresponses. We begin with a brief overview of a taxonomy for missing data and then address handling missing data in the IRT context.

In general, missing data (e.g., omitted responses) may be classified in terms of the mechanism that generated the missing values. According to Little and Rubin (1987), missing data may be classified as *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR; a.k.a., NMAR: not missing at random). MCAR refers to data in which the missing values are statistically independent of the values that could have been observed, as well as other variables. In contrast, when data are MAR, then the missing values are conditionally independent of one or more variable(s). In both of these cases, the data are missing at random either unconditionally (MCAR) or conditionally on one or more variables (MAR). If the data are neither MCAR nor MAR, then the missing values are considered to be MNAR and are *nonignorable*. Nonignorable missing values are data for which the probability of omission is related to what the response would be if the person had responded.

Various approaches for handling missing data have been developed. Some of these approaches share the goal of creating “complete data,” so standard analysis techniques may be applied. For instance, complete data may be created by deleting either the case that contains the missing value(s) either in its entirety or some subset of the case, or by replacing the missing value(s) by estimate(s) of what the missing value could have been. The replacement of the missing values by estimates is, in general, known as imputation. There are a number of single imputation methods (e.g., cold-deck imputation, hot-

deck imputation, mean substitution) as well as multiple imputation methods. Multiple imputation (MI) methods differ from single imputation methods by creating multiple (imputed) complete data sets to model the uncertainty in sampling from a population, whereas only one complete data set is created with single imputation. Other missing data methods are maximum likelihood-based. For greater detail, see C. H. Brown (1983); R. L. Brown (1994); Dillman, Eltinge, Groves, and Little (2002); Enders (2001, 2003); and Roth (1994).

Returning to the IRT context, there are various reasons why an individual's response vector may not contain responses to each item. We present three conditions that lead to missing data. The first condition is mentioned above. In the *missing by design* case (e.g., *not-presented* items), such as in adaptive testing (Appendix D) or simultaneous calibration (see Chapter 11), the nonresponses represent conditions in which the missingness process may be ignored for purposes of person location estimation (Mislevy & Wu, 1988, 1996). Therefore, the estimation is based only on the observed responses.

A second situation that produces missing data occurs when an individual has insufficient time to answer the item(s). These *not-reached* items are (typically) identified as collectively occurring at the end of an instrument (this assumes the individual responds to the test items in a serial fashion) and represent *speededness*. (Of course, the absence of not-reached items does not mean that speededness did not occur because respondents may randomly guess on items.) Although IRT should be applied to unspeeded tests, Lord (1980) stated that if we knew which items the examinee did not have time to consider, then these not-reached items may be ignored for person location estimation because they contain no readily quantifiable information about the individual's location (e.g., their proficiency). Therefore, when one has (some) missing data due to not-reached items, then the person's location is estimated using only the observed responses. However, this should not be interpreted as indicating that one should apply IRT to speeded instruments nor that these not-reached items are unaffected by being speeded. Speeded situations may lead to violation of the unidimensionality assumption and biased item parameter estimates. Research has shown that the speeded items' α_j s and δ_j s are overestimated and the χ_j s underestimated (Oshima, 1994). Because of the overestimation of α_j , the corresponding item information and, therefore, the instrument's total information becomes inflated. Identifying the speeded items as not-reached within BILOG mitigates the bias in item parameter estimation. (See Goegebeur, De Boeck, Wollack, and Cohen [2008] for a gradual process change model that models speededness as a person-specific effect.)

The third situation that produces missing data occurs when an examinee intentionally chooses to not respond to a question for which they do not know the answer. These *omitted responses* represent nonignorable missing data (Lord, 1980; Mislevy & Wu, 1988, 1996). Again, assuming that an individual responds in a serial fashion to an instrument, omitted responses may be distinguished from not-reached items because omits appear throughout the response vector and not just at the end of the vector. Lord (1980) has argued that omitted responses should not be ignored because an individual could obtain as high a proficiency estimate as they wished by simply answering only

those items they had confidence in answering correctly. This idea has found some support in Wang, Wainer, and Thissen's (1995) study on examinee item choice.

The effect of omitted responses on EAP person location estimates has been studied (de Ayala, Plake, & Impara, 2001; de Ayala, 2006; Finch, 2008; Glas & Pimentel, 2008; Rose, von Davier, & Xu, 2010). Results show that for dichotomous data, omits should not be treated as incorrect, nor should they be ignored; also see Lord (1974a, 1983c). However, using a fractional value of 0.5 in place of omitted values leads to improved person location estimation, compared with treating the omits as incorrect or using a fractional value equal to the reciprocal of the number of item options (i.e., $1/m$ where m is the number of response categories). (The $1/m$ approach assumes that an individual responds randomly to a multiple-choice item format and was suggested by Lord [1974a, 1980].) The results also seem to indicate that this would be true for MLE person location estimation. By using this fractional value, one is simply imputing a response for a binomial variable and thereby "smoothing" irregularities in the likelihood function. Although this research was conducted using the 3PL model, it appears that the results would apply to both the 1PL and 2PL models.

An alternative approach that may be fruitful in some situations is to treat omission as its own response category and apply a polytomous model such as the multiple-choice model or the nominal response model; both models are discussed in Chapter 9. Additionally, Holman and Glas (2005) present a "multiple" model approach that uses an IRT model to model the missing-data process and an IRT model for the observations. Missingness can also be addressed using MI. Several MI routines are available, including SAS proc mi, SPSS's multiple imputation (or EM from missing value analysis) (SPSS Incorporated, 2019), Missing Value Analysis (SYSTAT, 2017), or, for example, the R package mice (van Buuren & Groothuis-Oudshoorn, 2011, 2019). These routines assume that missing data are MAR. After imputation of omitted responses these complete data may then be calibrated.

The practitioner should be aware of several issues in the treatment of omits. For instance, in the context of proficiency assessment, all imputation procedures that produce complete data for analysis are, in effect, giving partial credit for an omitted response. For example, Lord's (1974a, 1980) suggested use of $1/m$ gives an individual partial credit worth, say 0.2 (i.e., $m = 5$), for having omitted an item. A second issue to be aware of is that using the same imputed value for all omits assumes that individuals located at different points can all be treated the same. These issues are raised so that the practitioner understands the assumptions that are being made with some of the missing data approaches discussed. However, these may or may not be of concern to a particular practitioner. For example, when IRT is used in personality testing or with attitude or interest inventories, these may be nonissues. A third issue to be noted is that omits tend to be associated with personality characteristics, demographic variables, and proficiency level (Mislevy & Wu, 1988; Stocking, Eignor, & Cook, 1988). Thus, in those situations where information on these variables is available, one may wish to use this information as covariates in the imputation process. Use of these covariate(s) may or may not have any meaningful impact on the person location estimates.

When calibrating a data set, it is good practice to identify items without responses

by some code. For instance, in the data file, not-reached items may be identified by a code of, say, 9, not-presented items by a code of 8, omitted items by a code of 7. With certain calibration programs (e.g., BILOG-MG), any ASCII character may be used (e.g., the letters “R” for not-reached, “P” for not-presented, and “O” for omit). In these cases, the code used must be identified for the program. With BILOG one would use the KFName, NFName, and/or OFName subcommands on the GLOBAL or INPUT command line, depending on the version of BILOG one is using. For BILOG, omitted responses must be identified as such, whereas with other programs any response code encountered in the data file that is not identified as a valid response is considered to reflect an omitted item. Omitted responses that have been identified by an omitted response code are, by default, treated as incorrect by BILOG.

Issues to Consider in Selecting among the 1PL, 2PL, and 3PL Models

The issues to be considered in selecting among the 1PL, 2PL, and 3PL models involve, in part, one’s philosophy of whether the data should fit the model or vice versa (see Chapter 2), as well as the application context (e.g., sample size, instrument characteristics and considerations, assumption tenability, political realities). Given that the 1PL model is the most restrictive of the three models, there have been a number of studies that have investigated use of the 1PL model when it misfits. For instance, Forsyth, Saissangjan, and Gilmer (1981) investigated the robustness of the Rasch model when the dimensionality and constant α assumptions are violated. Because their empirical data came from an examination using a multiple-choice item format, it was assumed that some examinees would engage in guessing. Forsyth et al. concluded that “the Rasch model does yield reasonably invariant item parameter and ability estimates . . . even though the assumptions of the model are not met” (p. 185). Similar results were obtained by Dinero and Haertel (1977) using simulation data.

Wainer and Wright (1980) stated, “It seems that the Rasch model yields rather good estimates of ability and difficulty even when its assumption of equal slopes is only roughly approximated” (p. 373). Furthermore, Lord and Novick (1968) stated, “It appears that if the number of items is very large, then inferences about an examinee’s ability based on his total test score will be very much the same whether” (p. 492) the Rasch model or the 3PL model is used. In this regard, recall that for the mathematics data example the Pearson correlation between the $\hat{\theta}$ s based on the 1PL and the 3PL models’ $\hat{\theta}$ s for the example’s data is 0.9764. For the other model combinations, we have a correlation of 0.9907 for the 1PL and the 2PL models’ $\hat{\theta}$ s, and for the 2PL and the 3PL models’ $\hat{\theta}$ s the correlation is 0.9859. Although these are all reasonably strong correlations, the correlations among the standard errors, $s_e(\hat{\theta})$ s, for the various model combinations paint a different picture. The correlation between the 1PL model estimated standard errors and those of the 2PL model is 0.9721, between the 1PL model and the 3PL model the correlation is 0.3318, and for the 2PL and the 3PL models’ estimated standard errors it is 0.2275. Therefore, in situations where confidence bands about $\hat{\theta}$ are used for

classification decisions, the same individual would be classified differently depending on the model used. Presumably, using longer instruments would allow for greater agreement among the standard errors. Moreover, the magnitude of the correlations between the 1PL, 2PL, and 3PL models' $\hat{\theta}$ s would be affected by the correlation between α_j and δ_j (Yen, 1981).

For samples of 200 or fewer, Lord (1983a) found that the Rasch model was slightly superior to the 2PL model in terms of person estimation. As previously mentioned, Thissen and Wainer (1982) studied the asymptotic standard errors of the one-, two-, and three-parameter models. They suggested fitting the 1PL model first and examining its model–data fit. If only a few items misfit and they could be omitted without adversely affecting the instrument (e.g., the validity of the $\hat{\theta}$ s), then one should consider removing them. However, if the omission of these misfitting items is problematic, then one should increase the sample size and try to fit the 2PL model (presumably the item[s] misfit is due to varying item discrimination). In contrast, Gustafsson (1980) suggested grouping the items into homogeneous subsets rather than removing them from the instrument. For instance, looking at the mathematics 2PL model calibration example, we see that in terms of the $\hat{\alpha}_j$ s there are three groupings of items. Items 3 and 4 are very similar in terms of their $\hat{\alpha}_j$ s, items 1 and 5 are somewhat similar to one another, and item 2 is substantially different from the other four items. Therefore, three subsets could be created for the mathematics data example. Assuming item misfit is due to varying item discrimination, we can alternatively use the OPLM model approach in which the item locations are estimated but the item discrimination(s) are imputed (Verhelst & Glas, 1995; Verhelst et al., 1995). The use of mixture models (see Appendix F, "Mixture Models"), as well as some of the models presented in von Davier and Carstensen (2007), may also provide additional solutions. (It should be recalled that the desirable properties of the Rasch model [e.g., specific objectivity] hold only when one has model–data fit.)

Yen (1981) advocates a process of first fitting all three models (i.e., 1PL, 2PL, 3PL) to the empirical data set of interest. Subsequently, simulation data sets are generated based on item parameter distributions that are similar to those found with the calibration of the empirical data set. For example, we would generate a data set using the 1PL model, another with the 2PL model, and so on. The final step involves comparing the fit analyses across models in conjunction with the fit analysis of the empirical data to facilitate model selection.

In a simulation study, Yen (1981) generated different data sets based on various models and compared the fit of the 1PL, 2PL, and 3PL models to these data. When she used the 3PL model for data generation, she found that the 2PL model fitted the data almost as well as the 3PL model did, although the item parameters estimates were not the same across the two models. She noted that when an item was difficult and had a moderate to high discrimination, it was difficult for the 2PL to model a nonzero lower asymptote. She concluded that although the 2PL model performed almost as well as the 3PL model in modeling the response vectors, one might observe sample dependency when difficult items have their discrimination parameters estimated with low-proficiency-level examinees.

As may be inferred from the above, there are variants of the dichotomous models.

For instance, it is possible to constrain the 3PL model to produce modified versions, such as constraining the α_j s to a constant value as well as the χ_j s to a nonzero value. This model is sometimes referred to as a modified 1PL model (Cressie & Holland, 1983; also see Kubinger & Draxler, 2007). Furthermore, one may use the 3PL model with the χ_j s for certain problematic items fixed to a constant nonzero value, whereas χ_j is estimated for other items. In general, for those situations where one is not holding χ_j s fixed, it would be prudent (as done above) to use a prior distribution on the χ_j s when estimating the lower asymptotes. In addition, with some data, one may obtain unreasonably large estimates of α_j (e.g., greater than 3). For these situations, use of a prior distribution on the α_j s may be in order.

As discussed in this chapter and the preceding chapters, it is the validity of the person location estimates that is paramount. From a pragmatic perspective, if convincing validity evidence can be accrued for person location estimates using a particular model in a particular application, then it would seem that the above arguments, though interesting in their own right, are somewhat irrelevant.¹⁸

Summary

The 3PL model attempts to obtain useful information from a response pattern over and above that contained in the response vector's observed score. To achieve this objective, the 3PL model consists of parameters that reflect the item's location and discrimination as well as the lower asymptote of the IRF. As is true with the 2PL model's IRFs, the 3PL model's IRFs may potentially cross because the 3PL model allows for varying discrimination. With the 3PL model, item discrimination is proportional to the slope of the IRF at the point of inflection and is equal to $0.25\alpha_j(1 - \chi_j)$. In addition, the 3PL model's IRFs may cross because the model allows for the lower asymptote parameters, χ_j s, to vary across items. The lower asymptote parameter is restricted to the range 0 to 1 (inclusive) and reflects the probability of obtaining a response of 1 by individuals who are extremely low on the latent variable continuum. The lower asymptote parameter is typically referred to as the pseudo-guessing parameter.

In previous chapters, fit is investigated in terms of item statistics, empirical and predicted IRFs, and examination of the invariance of item parameter estimates across random calibration subsamples. In this chapter, we also used ΔR_{Δ}^2 and ΔG^2 for assessing relative model–data fit. Moreover, we introduced an appropriateness index to gauge person fit and the Q_3^P and Q_3 statistics for assessing the tenability of the conditional independence assumption. The Q_3^P and Q_3 statistics may be useful for identifying sets of items that are exhibiting item dependence. When items are found to be interdependent, it may make sense to bundle them together and obtain an item score for the item parcel. The resulting item score is polytomous and ordinal in nature (i.e., a larger item score reflects more of the latent variable than does a smaller value). The analysis of these data can be accomplished through a polytomous model.

Chapter 7 introduces polytomous models that are derived from the Rasch model. These models, the partial credit and rating scale models, are appropriate for ordinal

polytomous data. These models assume that an instrument's items are equally effective in discriminating among individuals. As the models' names imply, the partial credit model can be used with data that reflect degrees of response correctness, whereas the rating scale model can be used with data from response formats, such as the Likert or summated ratings format. In actuality, both models are applicable to data that reflect degrees of response endorsement, but they differ from one another in their respective simplifying assumptions. In Chapter 8, use of polytomous models for ordinal data continues, but with models that are not based on the Rasch model.

Notes

1. Although the three-parameter model allows for the possibility that the lower asymptote is nonzero, the upper asymptote is still 1.0. That is, as θ approaches positive infinity, the probability of a response of 1 is 1.0 or, symbolically, $p(x=1|\theta \rightarrow \infty) \rightarrow 1$. An alternative model, the *four-parameter logistic* model (Barton & Lord, 1981), extends the three-parameter model to allow for the possibility that persons with very large θ s may still not have a success probability equal to 1 (see McDonald, 1967). The motivation behind the model's development was to improve person location estimation. For instance, if a person with a very large θ makes a clerical error on an easy item, then their estimate would be more drastically lowered using a model with an upper asymptote of 1 than when this asymptote was less than 1 (Barton & Lord, 1981). To address this situation, Barton and Lord (1981) introduced a parameter that reflected the IRF's upper asymptote (Υ_j) into the 3PL model. As a consequence, as θ goes to ∞ the probability of a response of 1 is Υ_j or, symbolically, $p(x=1|\theta \rightarrow \infty) \rightarrow \Upsilon_j$. The *four-parameter logistic* (4PL) model is

$$p(x_j = 1 | \theta, \alpha_j, \delta_j, \chi_j, \Upsilon_j) = \chi_j + (\Upsilon_j - \chi_j) \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} \quad (6.19)$$

Barton and Lord (1981) compared the model in Equation 6.19 to the 3PL model using empirical data. They found that the 3PL model did as well or better than the 4PL model. Barton and Lord concluded that "there is no compelling reason to urge the use of this <4PL> model" (p. 6). However, it should be noted that although the α_j s, δ_j s, and χ_j s were estimated (using JMLE), the Υ_j s were *not* estimated. Rather, the Υ_j s were held fixed at either 0.98 or 0.99. Given the study's design decisions, it is doubtful that this one study should be considered definitive. In contrast, Loken and Rulison (2010) using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2004) obtained promising estimation results in the estimation of all four item parameters. mirt and SAS proc irt can be used to estimate the 4PL model.

2. The first derivative of the 3PL model is

$$p'_j = \alpha_j(1-p_j) \frac{(p_j - \chi_j)}{(1-\chi_j)} ,$$

where

$$(1-p_j) = 1 - \left[\chi_j + (1-\chi_j) \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}} \right] = 1 - \left[\chi_j + \frac{(1-\chi_j)}{1+e^{-\alpha_j(\theta-\delta_j)}} \right].$$

Because by definition α_j is defined at $\theta = \delta_j$, p_j simplifies to

$$p_j = \chi_j + (1-\chi_j) \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}} = \chi_j + \frac{(1-\chi_j)}{1+e^0} = \chi_j + \frac{(1-\chi_j)}{2} = \frac{2\chi_j + 1 - \chi_j}{2} = \frac{1 + \chi_j}{2}$$

and $(1-p_j)$ simplifies to

$$(1-p_j) = 1 - \left[\chi_j + \frac{1-\chi_j}{1+e^{-\alpha_j(\theta-\delta_j)}} \right] = 1 - \frac{1+\chi_j}{2} = \frac{2-(1+\chi_j)}{2} = \frac{1-\chi_j}{2}.$$

By substitution for p_j in p'_j we obtain

$$p'_j = \alpha_j(1-p_j) \frac{(p_j - \chi_j)}{(1-\chi_j)} = \alpha_j \left[\frac{1-\chi_j}{2} \right] \left[\frac{\left(\frac{1+\chi_j}{2} - \chi_j \right)}{1-\chi_j} \right] = \alpha_j \left[\frac{\left(\frac{1+\chi_j}{2} - \chi_j \right)}{2} \right] = 0.25\alpha_j(1-\chi_j)$$

When the D scaling constant is used, then the slope for the 3PL model is

$$0.25D\alpha_j(1-\chi_j) = 0.425\alpha_j(1-\chi_j).$$

3. For example, assume we have a two-item instrument with $\alpha_1 = 2.0$, $\delta_1 = 0.0$, $\chi_1 = 0.25$ for the first item and $\alpha_2 = 1.0$, $\delta_2 = -0.5$, $\chi_2 = 0.0$ for the second item. According to the 3PL model, a person with the response vector $\underline{x} = 01$ will have an $\hat{\theta}$ of -0.55 . However, if we use the 2PL model (i.e., $\chi_1 = \chi_2 = 0.0$), then our $\hat{\theta}$ is -0.1558 . For the Rasch model (i.e., $\alpha_1 = \alpha_2 = 1.0$ and $\chi_1 = \chi_2 = 0.0$), our $\hat{\theta}$ is approximately -0.25 . Comparing these $\hat{\theta}$ s shows that one effect of including a nonzero χ_j in our model is to reduce the $\hat{\theta}$ s relative to not including χ_j .
4. Some users of the Rasch model have argued that the item discrimination parameter cannot be estimated as is done with the 2PL and 3PL models (e.g., see Wright, 1977b). According to Gustafsson (1980), when one has unequal discriminations, the item locations are related to the calibration sample's characteristics on the latent variable (e.g., a high- or low-proficiency group). In fact, he states that "it is difficult to make a distinction between the assumption of unidimensionality and the assumption of homogeneous item discrimination" (p. 208). Lumsden (1978) expresses a similar opinion: "Test scaling methods are self-contradictory if they assert both unidimensionality and different slopes for the ICC. . . . If the unidimensionality requirement is met, the Rasch (1960) one-parameter model will be realized" (p. 22). (Lumsden also suggested abandoning the two- and three-parameter normal ogives.) Gustafsson (1980) suggests that it may be prudent to investigate the robustness of the Rasch model in the face of varying item discriminations for specific applications.

5. According to Holland (1990a), there can be at most two parameters per item, and “models that contain three or more parameters per item can only estimate these parameters successfully for one of two reasons; either they are not applied to a large enough item set or the test is not unidimensional” (p. 17); also see Cressie and Holland (1983) and Holland (1990b). As such, there appear to be more parameters in the 3PL model than can be supported by a unidimensional test.
6. As is true with the two-parameter model, JMLE no longer seems to be used for parameter estimation with the three-parameter model. However, for completeness, we describe some of the past research in this area. The Hulin et al. (1982) study of JMLE presented in Chapter 5 also examined parameter estimation accuracy for two models (2PL, 3PL); this study had the additional factors of sample sizes (200, 500, 1000, 2000) and instrument length (15, 30, 60 items). They found that for a given condition the 2PL model results were better than those for the 3PL. However, for both models, and not surprisingly, the larger the sample size and the longer the instrument, the more accurate the estimates. In addition, the average error (i.e., root mean squared) in recovering the true IRFs for both models and using at least 30 items was no greater than 0.05 for a sample size of 1000 and less than 0.07 with 500 cases. In general, increasing the instrument’s length for a given sample size resulted in more accurate estimates.

Skaggs and Stevenson (1989) report a similar finding using LOGIST. They also found that the average error in recovering the true IRFs for the 15-item instrument was about 0.07, and for the 30-item length the average error was slightly below 0.055 when using a sample size of 500. These average errors decreased to about 0.05 and about 0.037 for the 15- and 30-item lengths, respectively, when the sample was quadrupled to 2000 cases. Lord (1968) suggests that the sample size be greater than 1000 and that instruments be at least 50 items long when using LOGIST. However, Swaminathan and Gifford (1983) found that reasonably good estimates can be obtained with a 1000-person sample and a 20-item instrument. Therefore, it appears that samples of 1000 or more with instruments of at least 20 items, and preferably longer, should be used with JMLE as implemented in LOGIST. However, work by Thissen and Wainer (1982) calls this sample size suggestion into question. For example, applying their observations to the 3PL model for an item with α_j of 1.5, $\delta_j = 2$ (or $\delta_j = -2$), and $\chi_j = 0.1$ would require 97,220, 22,142, and 46,743 individuals to estimate the item’s δ_j , α_j , and χ_j , respectively, with an accuracy of one-tenth. Therefore, the calibration sample size would be 97,220.

7. The testlet model is equivalent to a second-order model or a restricted bifactor model (Li, Bolt, & Fu, 2006; Rijmen, 2010).
8. Although the same calibration sample is used for the 1PL, 2PL, and 3PL model calibrations, the different models produced different estimates. The mean item location estimate for the 1PL, 2PL, and 3PL models are -0.403 , -0.400 , and 0.036 , respectively. Moreover, the mean discrimination estimate of 2.342 for the 3PL model is substantially greater than the common $\hat{\alpha} = 1.421$ found with the 1PL model or the 2PL model’s mean discrimination estimate of 1.459. This is due to the nonzero lower

asymptote as well as to differences in metrics. With respect to the former explanation, we see from a comparison of the 2PL and 3PL models' $\hat{\alpha}_j$ s that the 2PL model accommodates the nonzero asymptote by decreasing $\hat{\alpha}$ relative to what is obtained when we estimate the lower asymptote; for the 2PL model $\hat{\alpha}_1$ is 1.226 and for the 3PL model $\hat{\alpha}_1 = 1.921$. In fact, for all the items the 2PL model's $\hat{\alpha}_j$ s are less than the corresponding 3PL model's $\hat{\alpha}_j$ s. These lower 2PL model $\hat{\alpha}_j$ s are associated with a metric that, relative to the 3PL model's $\hat{\alpha}_j$ s, is stretched out and located lower than that of the 3PL model. In short, we have different metrics for the different model calibrations of the data. As such, the differences in the estimates across models for corresponding item parameters are partly due to a difference in metrics. Therefore, strictly speaking, we need to link the various metrics before directly comparing individual item parameter estimates across models.

9. The screening value of -0.2935 obtained in Appendix G, "Conditional Independence using Q_3^P ," can be used for evaluating Q_3^P . As mentioned in Appendix G, the generated data are conditionally independent. Theoretically, when the data are conditionally independent, there is no linearity in the residuals to partial out and the zero-order correlation Q_3 is equivalent to Q_3^P . Because our generated data contain a random error component, it is possible that the intercorrelations among one or more item pairs will not be equal to zero but will be very close to zero; sample size will also affect the equivalence of the partial and zero-order correlations. However, any difference from zero will not be meaningful and should not affect our conclusions. Consequently, we see that the item pairs identified using the gap approach (i.e., 2–3, 3–4, and 2–4) are also identified using the screening value.
10. A complementary approach for determining the df for evaluating ΔG^2 is to use the difference in the model's dfs . The df for a model is given by $2^L - (\text{number of item parameters}) - 1$, where L is the number of items on the instrument and the number of item parameters is based on the model and the number of items. For example, for the 3PL model there are three item parameters (α_j , δ_j , and χ_j), and for a, say, five-item instrument the number of items parameter is $3 \times 5 = 15$. Therefore, for the 3PL model the $df = 32 - 15 - 1 = 16$. For the 2PL model there are two item parameters (α_j and δ_j), and with a five-item instrument the $df = 32 - 10 - 1 = 21$. With the 1PL model, each item has a location (δ_j) and all items have a common α . Therefore, with a five-item instrument there are six parameters that are estimated and the model's $df = 32 - 6 - 1 = 25$. With BILOG, if one uses the keyword RASch, the program performs a 1PL estimation and then rescales the common α to be 1 and adjusts all the δ_j s accordingly; how this is done is demonstrated in Chapter 4. Therefore, with BILOG there are six, not five, item parameters estimated with the Rasch model. In contrast, a program like BIGSTEPS (or WINSTEPS) does not estimate a common α , and, as a result, there are only five δ_j s estimated; that is, the $df = 32 - 5 - 1 = 26$.
11. From Lord and Novick (1968) we have that the area under the item information function is

$$\int_{-\infty}^{\infty} I_j(\theta) = \alpha_j \frac{\chi_j \ln(\chi_j) + 1 - \chi_j}{1 - \chi_j}.$$

With the use of the D scaling constant in the 3PL model, the item information is

$$I_j(\theta) = \frac{D^2 \alpha^2 (1 - p_j)(p_j - \chi_j)^2}{(1 - \chi_j)^2 p_j} \quad (6.20)$$

and the corresponding area under the item information function is equal to

$$\int_{-\infty}^{\infty} I_j(\theta) = D \alpha_j \frac{\chi_j \ln(\chi_j) + 1 - \chi_j}{1 - \chi_j}.$$

12. To determine where an item has its maximum information, recall that α_j is proportional to the slope of the IRF at δ_j (i.e., the slope at δ_j is $0.25\alpha_j(1 - \chi_j)$). The offset from δ_j to where an item has its maximum information is obtained from the item information equation. By substitution into Equation 6.13 and rearranging terms, we have

$$I_j(\theta) = \alpha_j^2 \frac{e^{-\alpha_j(\theta - \delta_j)}}{1 + e^{-\alpha_j(\theta - \delta_j)}} (1 - \chi_j) \frac{e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}}{1 + e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}}.$$

Following Lord and Novick (1968) and maximizing $I_j(\theta)$ with respect to $\alpha_j(\theta - \delta_j)$ leads to

$$\begin{aligned} \frac{\partial}{\partial \alpha(\theta - \delta)} \ln I_j(\theta) &= \frac{\partial}{\partial \alpha(\theta - \delta)} \left[\ln \left(\frac{e^{-\alpha_j(\theta - \delta_j)}}{1 + e^{-\alpha_j(\theta - \delta_j)}} \right) + \ln \left(\frac{e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}}{1 + e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}} \right) \right] \\ &= 2 \left(\frac{e^{-\alpha_j(\theta - \delta_j)}}{1 + e^{-\alpha_j(\theta - \delta_j)}} \right) - 1 + \left(\frac{e^{-\alpha_j(\theta - \delta_j) + \ln(\chi_j)}}{1 + e^{-\alpha_j(\theta - \delta_j) - \ln(\chi_j)}} \right) \\ &= \frac{2}{1 + e^{-\alpha_j(\theta - \delta_j)}} - 1 + \frac{1}{1 + e^{-\alpha_j(\theta - \delta_j)/\chi_j}} \\ &= \frac{1 - e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} + \frac{\chi_j}{\chi_j + e^{\alpha_j(\theta - \delta_j)}} \\ &= \frac{2\chi_j + e^{\alpha_j(\theta - \delta_j)} - e^{2\alpha_j(\theta - \delta_j)}}{(\chi_j + e^{\alpha_j(\theta - \delta_j)})(1 + e^{\alpha_j(\theta - \delta_j)})}. \end{aligned}$$

To find the maximum of this last equation, its derivative is set to 0 and we solve for $\alpha_j(\theta - \delta_j)$

$$\frac{2\chi_j + e^{\alpha_j(\theta-\delta_j)} - e^{2\alpha_j(\theta-\delta_j)}}{(\chi_j + e^{\alpha_j(\theta-\delta_j)})(1 + e^{\alpha_j(\theta-\delta_j)})} = 0.$$

Because this equation is equal to 0.0, when its numerator equals zero we only need to be concerned with the numerator

$$2\chi_j + e^{\alpha_j(\theta-\delta_j)} - e^{2\alpha_j(\theta-\delta_j)} = 0.$$

This last equation is in the form of a quadratic (i.e., $f(x) = ax^2 + bx + c$, where a , b , and c are real constants and $x = e^t$; therefore, $2c + e^t + e^{2t}$). We can solve this last equation by using the quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

with $a = -1$, $b = 1$, and $c = 2\chi_j$. Because in this case, $a < 0$, we have two solutions: $1 + 4(2c) > 0$ and $-1/2(-1) = 0.5$. Using the quadratic formula, we obtain by substituting the values for a , b , and c

$$x = \frac{-1 \pm \sqrt{(-1)^2 - 4(-1)(2c)}}{2(-1)} = \frac{-1 \pm \sqrt{1+8c}}{-2}.$$

The solutions are

$$x = \frac{-1 + \sqrt{1+8c}}{-2} = \frac{1}{2} - \frac{\sqrt{1+8c}}{2} \text{ and } x = \frac{-1 - \sqrt{1+8c}}{-2} = \frac{1}{2} + \frac{\sqrt{1+8c}}{2}$$

We can eliminate

$$\frac{1}{2} - \frac{\sqrt{1+8c}}{2}$$

because it leads to having to take the log of a negative number. Therefore, we have $x = e^{\alpha_j(\theta-\delta_j)}$ and $\ln(x) = \alpha_j(\theta - \delta_j)$. By substitution

$$\ln\left(\frac{1}{2} + \frac{\sqrt{1+8c}}{2}\right) = \alpha_j(\theta - \delta_j) \frac{\ln\left(\frac{1+\sqrt{1+8c}}{2}\right)}{\alpha_j} = (\theta - \delta_j)$$

The item has the location of its maximum information at

$$\frac{\ln\left(\frac{1+\sqrt{1+8\chi_j}}{2}\right)}{\alpha_j} + \delta_j$$

and the offset is

$$\frac{\ln\left(\frac{1+\sqrt{1+8\chi_j}}{2}\right)}{\alpha_j}$$

That is, an item provides its maximum information at a location slightly higher than its δ_j . When $\chi_j = 0$, the offset equals 0.

13. The standard error for the person location estimate under the 3PL is

$$s_e(\hat{\theta}_i) = \sqrt{\sum_{j=1}^L \left[\frac{p_j(1-\chi_j)^2}{\alpha_j^2(1-p_j)p_j(1-\chi_j)^2} \right]} \quad (6.21)$$

where p_j is conditional on $\hat{\theta}_i$.

14. As mentioned in Chapter 5, the maximum information attainable by any scoring method is given by the total information function. Therefore, the optimal scoring weight for an item j is given by Equation 5.20

$$w_j(\theta) = \frac{p'_j}{p_j(1-p_j)}$$

Given that the first derivative for the 3PL model is

$$p'_j = \frac{\alpha_j(p_j - \chi_j)(1-p_j)}{(1-\chi_j)} \quad (6.22)$$

we have by substitution of Equation 6.22 into Equation 5.20 that the optimal scoring weight for the 3PL model is (Lord, 1980)

$$w_j(\theta) = \frac{\alpha_j(p_j - \chi_j)}{p_j(1-\chi_j)} = \frac{\alpha_j}{1 + \chi_j e^{-\alpha_j(\theta - \delta_j)}}. \quad (6.23)$$

Therefore, the optimal weight is a function of not only the item parameters, but also the person's location. As a result, with the 3PL model it is not possible to know the optimal scoring weight for an individual. Equation 6.23 shows that when $\chi_j = 0$, then $w_j(\theta) = \alpha_j$. Similarly, whenever θ is very large (i.e., $\theta \rightarrow \infty$), then the item's optimal weight approaches its discrimination (i.e., $w_j(\theta) \rightarrow \alpha_j$). In contrast, whenever θ is very small (i.e., $\theta \rightarrow -\infty$), then $p_j \rightarrow \chi_j$ and $w_j(\theta) \rightarrow 0$. In this latter condition, the respondent's location makes the item ineffective. With the scaling constant, D , Equation 6.23 becomes

$$w_j(\theta) = \frac{D\alpha_j}{1 + \chi_j e^{-D\alpha_j(\theta - \delta_j)}}$$

15. Equations 6.16–6.18 and the following equations are for maximum likelihood estimation. In addition to the information for each item parameter, we have information for the interrelationships among α_j , δ_j , and χ_j . Following Lord (1980) we have

$$I_{\alpha\delta_j} = \frac{\alpha_j}{(1-\chi_j)^2} \sum_i^N \left((\theta_i - \delta_j)(p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.24)$$

$$I_{\alpha\chi_j} = \frac{1}{(1-\chi_j)^2} \sum_i^N \left((\theta_i - \delta_j)(p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.25)$$

and

$$I_{\delta\chi_j} = \frac{\alpha_j}{(1-\chi_j)^2} \sum_i^N \left((p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right). \quad (6.26)$$

Collectively, Equations 6.16–6.18 and 6.24–6.26 form the information matrix (\mathbf{I}_j) for item j

$$\mathbf{I}_j = \begin{bmatrix} \text{Eq. 6.16} \\ \text{Eq. 6.24} & \text{Eq. 6.17} \\ \text{Eq. 6.25} & \text{Eq. 6.26} & \text{Eq. 6.18} \end{bmatrix}. \quad (6.27)$$

The reciprocals of the square root of the main diagonal elements are the estimates of the standard errors of α_j , δ_j , and χ_j . On the normal metric, the corresponding item parameter information formulas are (Lord, 1980)

$$I_{\alpha_j} = \frac{D^2}{(1-\chi_j)^2} \sum_i^N \left((\theta_i - \delta_j)^2 (p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.28)$$

$$I_{\delta_j} = \frac{D^2 \alpha_j^2}{(1-\chi_j)^2} \sum_i^N \left((p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.29)$$

$$I_{\chi_j} = \frac{1}{(1-\chi_j)^2} \sum_i^N \frac{(1-p_j)}{p_j}, \quad (6.30)$$

$$I_{\alpha\delta_j} = \frac{D^2 \alpha_j}{(1-\chi_j)^2} \sum_i^N \left((\theta_i - \delta_j)(p_j - \chi_j)^2 \frac{(1-p_j)}{p_j} \right), \quad (6.31)$$

$$I_{\alpha\chi_j} = \frac{D}{(1-\chi_j)^2} \sum_i^N \left((\theta_i - \delta_j)(p_j - \chi_j) \frac{(1-p_j)}{p_j} \right), \quad (6.32)$$

and

$$I_{\delta\chi_j} = -\frac{D \alpha_j}{(1-\chi_j)^2} \sum_i^N \left((p_j - \chi_j) \frac{(1-p_j)}{p_j} \right). \quad (6.33)$$

16. For completeness, the information functions for estimating α_j for all five items are shown in Figure 6.18. As can be seen, the bimodal pattern exhibited in Figure 6.14 is true for all items. The different modal values reflect the magnitude of the nonzero

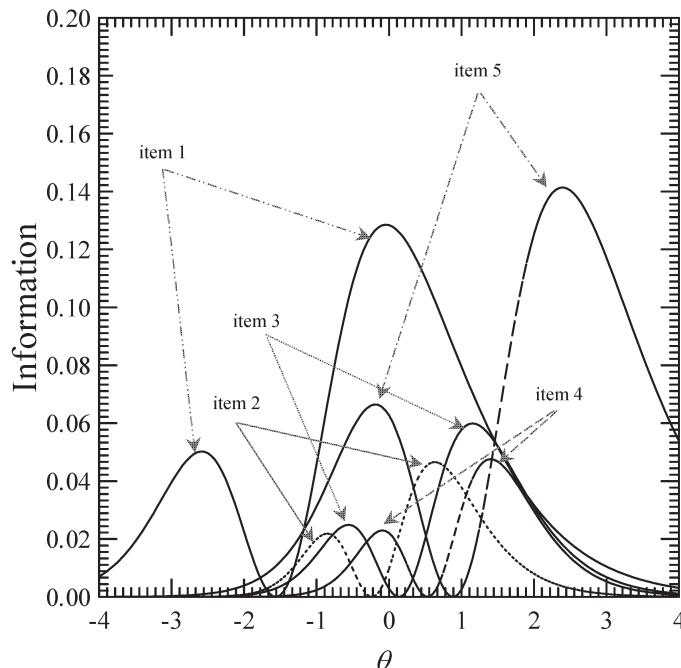


FIGURE 6.18. Information for estimating α_j as a function of θ for each of the five math items.

χ_j s. It is also apparent that the modes occur on opposite sides of the item's location, with the leftmost mode always less than the rightmost mode. This characteristic is a reflection of positive α_j s (i.e., if the α_j s are negative, then the leftmost mode would be greater than the rightmost mode). The location of the minimum of the information function between the two modes corresponds to the item's δ_j ; this minimum information is 0.

17. Typically, the TCC is depicted as ogival shaped and as resembling an IRF. However, the TCC's shape is a function of not only the number of items, but also the calibration model and the distribution/characteristics of the item parameter estimates. For example, if our $\hat{\delta}_j$ s are more widely spaced than those used in Figure 6.17, the TCC's shape would change. Figure 6.19 contains the TCC for a five-item set that uses the same $\hat{\alpha}_j$ s and $\hat{\chi}_j$ s as in Figure 6.17, but with $\hat{\delta}_1 = -3.0$, $\hat{\delta}_2 = -2$, $\hat{\delta}_3 = 0.0$, $\hat{\delta}_4 = 2$, and $\hat{\delta}_5 = 3.0$. Clearly, this TCC is still monotonically nondecreasing, but it also contains ridges. (One needs to extend the abscissa to see that the TCC is asymptotic with $\sum \chi_j$.)
18. Based on the work of Yen (1981), it appears that whenever one applies an inappropriate model to a data set, one may obtain *sample-dependent* estimates (i.e., a contradiction to one of IRT's potential advantages). Therefore, adopting a model that expresses one's intentions and does not simply describe the data appears to be a prudent strategy. From a philosophical perspective, because all models are false, then this begs the question as to whether one may obtain sample-independent estimates

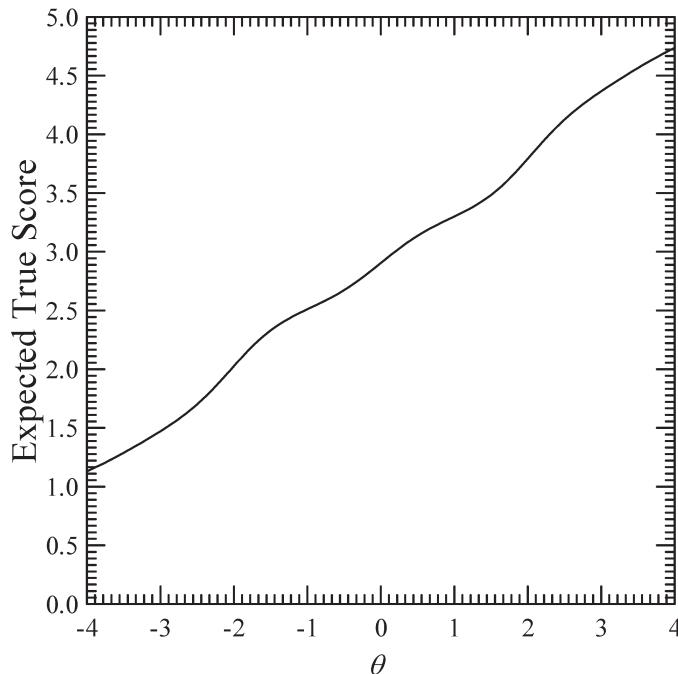


FIGURE 6.19. TCC for widely spaced δ s.

in any truly absolute fashion. It is conjectured that, most likely, the best that one may be able to achieve is sample-independent estimates for a particular range of data (as demonstrated in Chapter 3). If these data represent the situations in which one is primarily interested, then whether one may obtain sample independent estimates in an absolute fashion may be academic.

7

Rasch Models for Ordered Polytomous Data

The models discussed so far are appropriate for dichotomous response data. This type of data can be obtained either directly by using a two-option item (e.g., the true/false item format used with the MMPI [Swenson, Pearson, & Osborne, 1973]) or indirectly by recoding an individual's response as 0 or 1 through a scoring paradigm or by passing judgment on the individual's response. The response of 0 reflects the absence of some characteristic or, in the case of proficiency testing, an incorrect response. Conversely, the response of 1 represents the presence of the characteristic, a respondent's endorsement, or in the case of proficiency testing, a correct response. However, in some situations, our data may have more than two response categories. For instance, judges may rate a person's performance on a rating scale, a Likert scale allows varying degrees of agreement to an item, or we can assign codes that reflect an answer's degree of correctness.

In these examples, our data are polytomous (i.e., more than two response categories) and inherently ordered. By "ordered" we mean that some responses indicate more (or less) of what is being measured than other responses. For example, assume that we allow a person to respond to the statement "The Congress should legislate that English be the national language of the United States" by using one of four possible response categories, such as "strongly agree," "agree," "disagree," "strongly disagree." These response categories, and therefore the person's response, are ordered on a scale that represents a favorable to an unfavorable attitude. As another example, assume that the amount of credit an answer is given is directly related to the degree of correctness of the response. As a result, higher credit indicates greater, say mathematics proficiency, than does lower credit.

Below we discuss models derived from the Rasch model that are appropriate for the above examples of ordered polytomous data. In the subsequent chapter we present two additional models, the generalized partial credit and graded response models, which may also be used for analysis of ordered polytomous data. Models for unordered polytomous data are discussed in Chapter 9.

Conceptual Development of the Partial Credit Model

The example data used in previous chapters contain examinee responses to a mathematics examination. The test administrator took the examinees' observed responses and classified them into one of two categories, correct or incorrect. This dichotomization treats all the incorrect answers as equivalent to one another. As such, all the mathematical operations required to answer an item are considered *una voce* (i.e., as a "single operation"). If an examinee correctly performs all the operations, then they have performed "one correct operation" and the response is categorized in the correct category with an assigned value of 1. Otherwise, their response is categorized as incorrect and it receives a value of 0. Therefore, the value assigned to the examinee responses reflects *only* whether the examinee has correctly performed (heuristically) one "operation." In contrast, consider the item

$$(6/3) + 2 = ?$$

A scoring rubric for this item might be based on the assumptions that to correctly answer this item certain operations must be performed correctly and that it is not possible to correctly guess on these operations. The first operation is the evaluation of the quotient 6/3, and the second operation is the addition of the numeral 2 to the first operation's result. Zero points would be assigned for incorrectly performing the first operation. Because this item consists of two operations, we can assign partial credit (e.g., 1 point) for correctly performing only the first operation and full credit for correctly performing the first and second operations (i.e., the two points reflects the number of correctly performed operations). We believe that if some credit is assigned for partially correct responses, the partially correct responses can provide useful information for estimating a person's location.

Implied in this rubric is that one cannot obtain one point for correctly performing *only* the second operation. That is, the only way to obtain one point is by successfully performing only the first operation. Therefore, for this item j there are three possible integer scores (x_j) of 0, 1, 2 (i.e., $x_j = \{0, 1, 2\}$). These scores are called *category scores* and indicate the number of or a count of the successfully performed operations. As a result, higher-category scores indicate a higher level of overall performance than do lower-category scores. In this approach, and in general, the examinees' responses are categorized into $m_j + 1$ scores (i.e., $x_j = \{0, 1, \dots, m_j\}$), where m_j is the number of "operations" to correctly answer item j . For the above example item $m_j = 2$.

One approach for modeling ordered polytomous data involves decomposing the responses into a series of ordered pairs of adjacent categories or category scores and then successively "applying" a dichotomous model to each pair. Masters (1982) used this approach in developing his partial credit model, which is described below.

Assume there is a point on the latent variable continuum below which an individual provides a particular response (e.g., $x_j = 0$) and above which the person provides the next higher response (e.g., $x_j = 1$). As such, this point indicates the transition from one category score to the next category score. In the current context, a polytomously scored

item has multiple ordered response categories (or category scores) with adjacent categories separated by such a transition point. Let each transition point be indexed by h and the *location* of each transition point h on the continuum for item j be represented by δ_{jh} (i.e., the *transition location parameter*). For the example item these two transition points would be symbolized as δ_{j1} and δ_{j2} ; they are interpreted below.

Generally speaking, the Rasch model specifies the probability, p_j , of the occurrence of an event b , and $(1 - p_j)$ specifies the probability of b 's complementary event \bar{b} (e.g., $\bar{b} = 1$ and $\bar{b} = 0$); the events b and \bar{b} are mutually exclusive and jointly exhaustive. (Sometimes the events b and \bar{b} are referred to as success and “not success,” respectively.) When we apply the Rasch model to ordered polytomous data, the events are the adjacent category scores or the adjacent response categories. For instance, for the example item, “ $(6/3) + 2 = ?$,” we have two pairs of adjacent category scores. Pair one consists of $x_j = \{0, 1\}$ and the second pair is $x_j = \{1, 2\}$. Each pair of adjacent category scores has a transition point. The first transition point (δ_{j1}) reflects the shift in pair one from $x_j = 0$ to $x_j = 1$, whereas the second transition point (δ_{j2}) is for the progression in pair two from $x_j = 1$ to $x_j = 2$. In terms of the events terminology, for pair one the event \bar{b} reflects $x_j = 0$ and the event b represents $x_j = 1$ (i.e., b reflects a successful transition from 0 to 1). For the second pair, the event \bar{b} reflects $x_j = 1$ and the event b represents $x_j = 2$ (i.e., b represents a successful transition from 1 to 2). The pair's corresponding (transition) location parameter can be used in the Rasch model to calculate the probability of event b .

Because the Rasch model can be used to calculate the probability of the successful transition from one category score to the next category score, its complement specifies the probability of an individual who is *not* successful in the transition from one category score to the next category score. (For presentation ease, we drop the subscript j and retain the subscript h on δ in the following.) For example, if an individual is *not* successful in the transition from $x_j = 0$ to $x_j = 1$, then the corresponding probability according to the Rasch model is

$$p(x_j = 0) = 1 - p(x_j = 1) = 1 - \frac{e^{(\theta-\delta_1)}}{1+e^{(\theta-\delta_1)}} = \frac{1}{1+e^{(\theta-\delta_1)}} = \frac{e^0}{e^0 + e^{(\theta-\delta_1)}}.$$

More generally, we may write

$$p(x_j = 0) = \frac{e^0}{\varphi},$$

where φ reflects all the possible outcomes. In contrast, if the individual is successful in the transition from $x_j = 0$ to $x_j = 1$ (i.e., pair one from above), then based on the distance between the person's location and the (transition) location parameter, we have

$$p(x_j = 1 | x_j = 0) = \frac{e^{(\theta-\delta_1)}}{\varphi}.$$

This may be interpreted as the probability of observing a category score of 1 *rather* than a category score of 0. Similarly, for pair two we have the probability of observing a category score of 2 *rather* than a category score of 1 as

$$p(x_j = 2 | x_j = 1) = \frac{e^{(\theta - \delta_2)}}{\varphi}.$$

In the preceding, we have applied the Rasch model to the separate dichotomizations but have not considered the ordinal relationship between the three possible outcomes. These separate calculations may be “aggregated” by invoking this ordinal relationship. In the following, we temporarily omit displaying the denominator, φ ; however, the appropriate denominator is implied for the statements that read “the probability of . . .”

If an individual fails to make the transition from $x_j = 0$ to $x_j = 1$, then the probability of this event is e^0 . However, for an individual to obtain an $x_j = 1$, then this person needs to “pass through” a response of 0 and the first transition point, δ_{j1} . The probability of both of these is given by adding the (mutually exclusive) events of 0 and 1, that is, $e^{0+(\theta - \delta_1)}$.

In a similar fashion, for a person to obtain an $x_j = 2$, then this person needs to “pass through” the second transition point, δ_{j2} . However, to get to δ_{j2} , they would have to pass through a category score of 1 (i.e., the first transition point); this idea is embodied in the phrasing above, “category score of 2 rather than a category score of 1.” To pass through δ_{j1} , the person would have to pass through a response of 0. Therefore, to obtain an $x_j = 2$, the individual passes through $x_j = 0$ (i.e., e^0), passes through $x_j = 1$ (i.e., $e^{0+(\theta - \delta_1)}$), and then passes through the second transition point (i.e., $e^{0+(\theta - \delta_1) + (\theta - \delta_2)}$). As a consequence, the probability of $x_j = 2$ is given by $e^{0+(\theta - \delta_1) + (\theta - \delta_2)}$. Figure 7.1 contains a schematic representation of these sequences where (a), (b), and (c) conceptually reflect the processes for attaining the category scores of 0, 1, and 2, respectively. As can be seen, the category score determines the number of aggregated distances between a person’s location and the (transition) location parameter(s).

When each of the three terms (i.e., e^0 , $e^{0+(\theta - \delta_1)}$, $e^{0+(\theta - \delta_1) + (\theta - \delta_2)}$) is divided by φ , one obtains the probability of the category scores of 0, 1, and 2, respectively. Consistent with the definition of a probability, the denominator φ is the sum of the three mutually exclusive outcomes.

A general expression that incorporates the principles just outlined is given by the *partial credit* (PC) model (Masters, 1982). The PC model specifies the conditional probability that a randomly selected examinee with latent location θ obtains a category score of x_j is

$$p(x_j | \theta, \delta_{jh}) = \frac{\exp\left[\sum_{h=0}^{x_j} (\theta - \delta_{jh})\right]}{e^0 + \sum_{k=1}^{m_j} \exp\left[\sum_{h=0}^k (\theta - \delta_{jh})\right]} = \frac{\exp\left[\sum_{h=0}^{x_j} (\theta - \delta_{jh})\right]}{\sum_{k=0}^{m_j} \exp\left[\sum_{h=0}^k (\theta - \delta_{jh})\right]}. \quad (7.1)$$

For convenience

$$\sum_{h=0}^k (\theta - \delta_{jh}) \equiv 0,$$

and for notational ease “ $\exp[z]$ ” is used in lieu of “ e^z .” The transition location parameter for item j , δ_{jh} , is sometimes referred to as a “step difficulty” parameter or “step param-

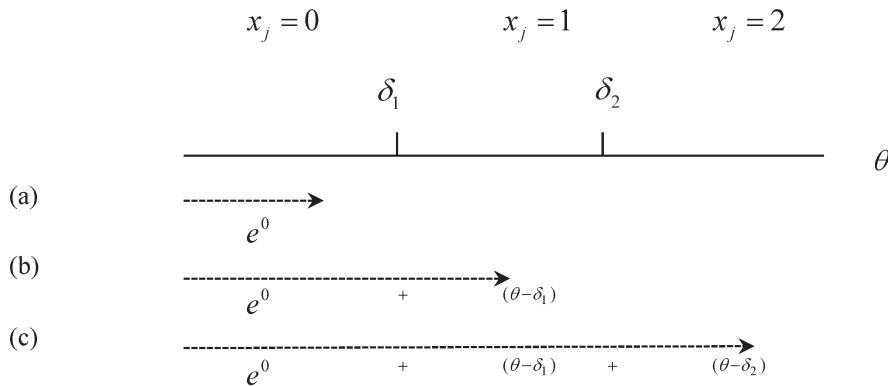


FIGURE 7.1. Schematic representation of the processes for obtaining the category scores of 0, 1, and 2.

eter.” Alternatively, we can view δ_{jh} as a function of an overall item location, δ_j , plus an offset or deviation from this location, τ_{jh} (i.e., $\delta_{jh} = \delta_j + \tau_{jh}$); $\sum_{h=1}^{m_j} \tau_{jh} = 0$. As mentioned above, we call δ_{jh} the *transition location parameter*.¹ In effect, δ_{jh} reflects the relative difficulty in endorsing category h over category $(h - 1)$. The use of a subscript on m (i.e., m_j) indicates that the number of category scores may vary across items. Accordingly, the PC model may be applied to items that are polytomously scored with a varying number of category scores, are dichotomously scored, or consist of both dichotomous and polytomously scored items.² In the following, p_{xj} is used for $p(x_j | \theta, \delta_{jh})$.

Although for the PC model the response category scores must be ordered this does not mean that the transition locations are necessarily ordered. This is easiest to see in the context of proficiency assessment. For instance, it is obvious for the example item above that the second operation (i.e., the addition of 2) is a comparatively easier operation than is the first operation (i.e., the division of 6 by 3). Therefore, the corresponding δ_{jh} s are not in increasing order. Moreover, it is important to note that the δ_{jh} s are conditionally defined and cannot be interpreted simply as independent pieces of an item. Masters (1988) states that the δ_{jh} s “are more appropriately interpreted as a set of parameters associated with item $\langle j \rangle$, none of which can be interpreted meaningfully outside the context of the entire item in which it occurs” (p. 23).

As an extension of the Rasch model, the PC model is predicated on a unidimensional construct of interest and on the notion that items discriminate among respondents to the same degree. The next chapter presents an alternative model, the generalized partial credit model, that relaxes the equal discrimination assumption.

The probability of obtaining a particular category score as a function of θ may be graphically represented in an *option response function* (ORF); ORFs are sometimes referred to as *category probability curves*, *category response functions*, *operating characteristic curves*, or *option characteristic curves*, or, generically, *trace lines*. ORFs may be viewed as an extension of the dichotomous models’ IRFs to polytomous data.

Figure 7.2 contains the ORFs for an item j worth two points ($m_j = 2$) with ordered transition locations. The first transition location occurs at -1.0 (i.e., $\delta_{j1} = -1.0$) and the

second transition location at 1.0 (i.e., $\delta_{j2} = 1.0$). An example of an item that would produce ordered transition locations is “ $(8 + 4)/3 = ?$ ” For this item the first operation is the sum of 8 and 4, and the second, more difficult, operation is the division of this sum by 3; $m_j = 2$. The proficiency required to correctly perform the first operation and obtain a category score of 1 ($x_j = 1$) is less than that required to correctly perform the second operation to obtain a category score of 2 ($x_j = 2$). As a result, the corresponding transition locations would reflect this ordering of proficiencies.

As Figure 7.2 shows, the transition location parameter is the point of transition from one category (score) to the next. Stated another way, the transition location is the point where the probability of responding in two adjacent categories is equal. This reflects that an individual's response mechanism is really a binary “choice” between two adjacent categories. That is, the probability of selecting a particular response category over the previous one is governed by the dichotomous Rasch model. For example, the transition location at -1.0 is the transition between a category score of 0 and that of 1, whereas the transition location at 1.0 represents the transition between the (partial credit) score of 1 to a (full credit) score of 2. As one moves in either direction away from a transition location, the probability of obtaining one particular category score increases, while the probability of obtaining the other category scores decreases.

One may interpret these ORFs as indicating that individuals located below -1.0 are most likely to obtain a category score of 0 (i.e., $x_j = 0$). However, some of the examinees located below -1.0 may obtain a score of 1 (i.e., $x_j = 1$ line segment below -1.0), albeit with a lower probability than a score of 0. Moreover, there is a substantially smaller probability that individuals located below -1.0 will correctly respond to this item (i.e., a category score of 2; the portion of the $x_j = 2$ line below -1.0). Persons located between -1.0 and 1.0 are most likely to obtain a category score of 1 (e.g., correctly performing one operation), and those located above 1.0 are increasingly likely to obtain a category score of 2 (e.g., correctly performing both operations). In each of these latter cases, there is some nonzero probability that persons will not respond in the most likely category. The sum of the probabilities across category scores for a given θ is 1.0 .³

In contrast to the ordered transition locations seen in Figure 7.2, Figure 7.3 contains the ORFs for a two-point item in which the transition locations are in reverse order. (When adjacent transition locations are in decreasing order, we call it a “reversal” [e.g., see Dodd, 1984].) For this item, the transition from no credit (i.e., a category score of 0) to that of partial credit (1 point) occurs at 1.0, and the transition from partial credit to full credit (2 points) occurs at -1.0 . If this is a proficiency item, then the first transition from a category score of 0 to that of 1 would be substantially more difficult than the second transition from a category score of 1 to that of 2 because the first transition is located to the right of the second transition location. The item “ $(6/3) + 2 = ?$ ” is an example of this situation (i.e., the first operation, $(6/3)$, is more difficult than the second operation, the addition of 2).

Because of the transition location reversal, this item is functioning similarly to a dichotomous item. Below $\theta = 0.0$, persons are most likely to provide a category score of 0 (e.g., obtain no credit on this item; $x_j = 0$ line), and individuals above this point are most likely to respond with a category score of 2 (e.g., obtain full credit on this item; $x_j = 2$

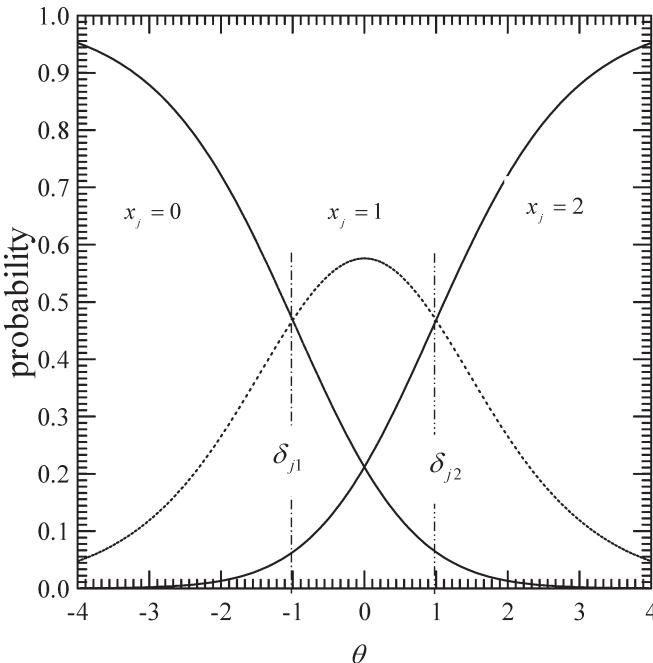


FIGURE 7.2. PC model ORFs for a two-point item j with $\alpha = -1.0$ and $\beta = 1.0$.

line). Comparatively speaking, the probability that persons will obtain a category score of 1 (e.g., obtain partial credit on this item; $x_j = 1$ line) is less than that for the other two possibilities. Intuitively, this makes some sense. That is, if one has the capability to correctly perform the more difficult first operation on this item, then it is very likely that one would successfully perform the second and easier operation.

The ORFs for any item always consist of, at a minimum, one monotonically nondecreasing ORF and one monotonically nonincreasing ORF. With dichotomous data, the monotonically nondecreasing ORF represents a category score of 1 and has a positive α (i.e., this ORF is the IRF). The monotonically nonincreasing ORF reflects a category score of 0 and has the negative of α . Because the two ORFs intersect at the item's location, we can consider δ to be the threshold or transition location between a response of 0 and a response of 1. For items with more than two response categories, we potentially have one unimodal ORF for each additional response category score.

Conceptual Parameter Estimation of the PC Model

The general principles presented in Chapters 2 through 4 as well as Appendices A and B are applicable for estimating the multiple item and person parameters. In general, one obtains the likelihood function for the response data. The person and item values that maximize the likelihood function are taken as the estimates of the person and item locations. With the PC model, this estimation capitalizes on the existence of sufficient

statistics. With respect to persons, the observed score is a sufficient statistic for estimating a person's location. Therefore, individuals who obtain the same observed score (i.e., the simple unweighted sum of their item responses) obtain the same location estimate. Thus, an individual's pattern of responses does not provide any more information for estimating the person's location than does their observed score. In regard to items, it is the simple count of the number of respondents for each category score, not the actual response pattern across individuals, that contains all the information required for estimating the transition locations. Stated another way, the counts are sufficient statistics for estimating the transition locations for an item. Therefore, item categories that have the same count have the same transition location estimate. The details of the estimation for the PC model may be found in Wright and Masters (1982) and Masters (1982); also see Wilson and Adams (1993).⁴

Example: Application of the PC Model to a Reasoning Ability Instrument, MMLE, flexMIRT

Assume the data have been cleaned and that we have evidence supporting our unidimensional model as a reasonable representation of the data. In addition, assume that we have validity evidence supporting the use of our instrument. Although we could use BIGSTEPS/WINSTEPS (or mixRasch) for performing our calibration, we use flexMIRT

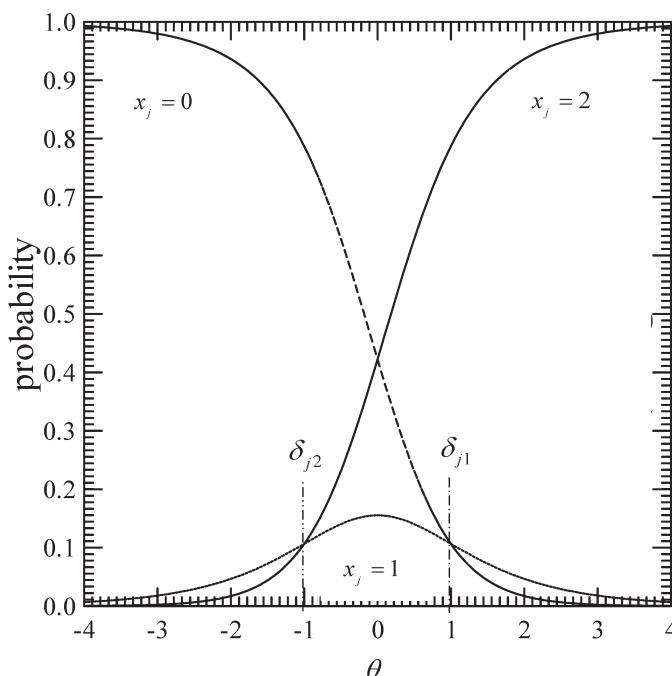


FIGURE 7.3. PC model ORFs for a two-point item j with $\gamma = 1.0$ and -1.0 .

for our example; subsequently, we use `mirt`.⁵ flexMIRT implements MMLE for parameter estimation for the dichotomous models discussed above as well as for polytomous models such as the graded response, generalized partial credit, and nominal response models, among others; these polytomous models are discussed in Chapters 8 and 9. (flexMIRT also estimates cognitive diagnostic models and uses the Metropolis-Hastings Robbins-Monro estimation algorithm for multidimensional models.)

Our example data concern reasoning ability and come from the General Social Science survey (National Opinion Research Center, 2003). The instrument consists of a series of questions that ask how two things are alike. For instance, “In what way are a dog and a lion alike?” (item 2), “In what way are an egg and a seed alike?” (item 4), or “In what way are a table and a chair alike?” (item 5). The responses were graded using three categories (incorrect—0 points, partially correct—1 point, and correct—2 points) according to the quality of the response. Specifically, the rubric emphasized the response’s degree of abstraction in determining the response score. Two points were awarded for pertinent general categorizations; one point for the naming of one or more common properties or functions of the members of a pair (this is considered a more concrete problem-solving approach); and 0 points for the naming of specific properties of each member of the pair, generalizations that were incorrect or not pertinent, differences between the members of the pair, or clearly wrong responses. For example, to the question “In what way are an orange and a banana alike?” (item 1) 2 points would be awarded if the response is “both are fruit,” 1 point for a response such as “both are food,” and 0 points for a response such as “both are round.” There are eight such items on the instrument.

flexMIRT can be used to estimate only item parameters, only person locations from previously obtained item parameters, or both in a single execution. Because flexMIRT does not use formatted reads the data must be space, comma, or tab delimited, and the file cannot have a header line (e.g., containing variable names). Additionally, nonzero-based responses (e.g., 12212) vectors need to be recoded to be zero-based (e.g., 01101); this can be done within flexMIRT (see Endnote 19). The command file consists of Project, Options, Groups, and Constraints sections. Within each of these required sections, we specify an analysis title (Project), the analysis to be performed (Options), data file information (Groups), and any constraints to be imposed (Constraints). It is through this last section that one obtains estimates for non-primary models (e.g., the 1PL model). The primary models have a Model option statement such as ThreePL, Graded, Nominal, GPC for 3PL, graded response, nominal response, and the generalized partial credit model, respectively. The non-primary models are implemented as special cases of the primary models. For instance, by imposing constraints on the 3PL model, one can obtain estimates for the 1PL/Rasch and 2PL models. Comments can be embedded into the command file by using two forward slashes (“//”) or, as is done in SAS, using a forward slash-asterisk (“/*”) and asterisk-forward slash (“*/”) to mark the beginning and end of a comment, respectively. All command lines end are terminated by a semicolon and contain one or more options.⁶

The command file for the PC calibration of the reasoning ability data is shown in Table 7.1; our file is called `PCCalib` with the default extension `flexmirt`. Briefly,

the <Project> section contains two required commands or keywords (Title, Description) that one can use to annotate the output with the corresponding text within single or double quotes; the quotes must be present even if no text is provided (e.g., Description = "");.

In the <Options> section, we specify (1) the type of analysis to perform (Mode = Calibration), (2) the extent of model and item-level fit information (GOF = Extended; M2 = Full), (3) estimating person locations using EAP (SCORE = EAP) and to save them to an external file (saveSCO = Yes), as well as (4) to calculate the information functions using 81 equal intervals from -4 to 4 (FisherInf = 81,4.0) and to save this information to an external file (saveINF = Yes); our interval width is $(|-4|+|4|)/81 = 0.1$. We also save the item parameter estimates (savePRM = Yes).

In the <Groups> section, we label our single-group analysis as OnlyGroup, provide our data filename (File = "ALIKE.DAT"), as well as information about the number and names of the items (Varnames = i1-i8) and their respective response categories (Ncats(i1-i8) = 3). Unless otherwise specified on the FILE command, the data file is assumed to reside in the same subdirectory (a.k.a., folder) as the command file.

To perform the PC model calibration, we use the <Groups> and <Constraints> sections. In the <Groups> section, we specify the generalized partial credit (GPC) model (Chapter 8) should be used for all eight items (Model(i1-i8) = GPC(3)). In the <Constraints> section, we fix the starting value of α to 1.0 (VALUE (i1-i8), slope, 1.0) and specify that item discrimination (FIX (i1-i8), slope) not be estimated.⁷

Table 7.2 contains our output. This output is automatically saved to an ASCII file with the extension 'txt' and '-irt' appended to the command filename; an existing file with the same name is automatically replaced. For instance, our output file is PCCalib-irt.txt. A similar approach is used for naming any file to be created and saved (e.g., saveSCO). The output file as well as any other files' content is accessible from the appropriately labeled tab in the output window.

The output begins with an echoing of the calibration control parameters and a specification of number of items (Number of Items), number of examinees (Number of Cases), and dimensionality (Latent Dimensions). For example, the program read Number of Cases is 2942 cases, there are eight items, and so on. Following this section, a table containing the number of response categories for each item follows (Categories column). The Model, Ta, and Tc portion shows that the GPC model is implemented as a special case of the nominal model with transformation matrices Ta and Tc that are using trend contrasts.⁸

The calibration control parameters show that the maximum number of EM cycles is 500 (i.e., Maximum number of cycles: 500) and the maximum number of M-steps is 100 (i.e., Maximum number of M-step iterations: 100). For the EM cycles and M-steps phases, the convergence criteria are 0.0001 and 0.0000001, respectively. We see that our SE = Fisher command for how to calculate our standard errors is echoed as Standard error computation algorithm: Fisher (Expected). Our specified starting values are displayed (User-Defined Initial Values); P #

TABLE 7.1. Command File for the flexMIRT PC Model Calibration Example

```

<Project>
  Title = "PC calibration, ALIKE data - slope= 1";
  Description = "Item & person parameter estimates - single run ";

<Options>
  Mode = Calibration;
  GOF = Extended;
  M2=Full;                                // M2 stat full marginal, not 1st- & 2nd-order subtables
  FitNullModel=Yes;                         // obtain TLI/NNFI
  NumDec = 3;                               // decimals for item parameter estimates
  savePRM= Yes;                            // save item parameter estimates
  SCORE= EAP;                              // calculate EAP person location estimates
  saveSCO= Yes;                            // calculate & save person location estimates
  FisherInf=81,4.0;                        // 81 equal intervals from -4 to 4 to calculate info
  saveINF= Yes;                            // save information function - items/scale
  SE= Fisher;                             // use Fisher info matrix to calculate standard errors

<Groups>
  %OnlyGroup%
  File = "ALIKE.DAT";      // space delimited file
  Varnames = i1-i8;
  N = 2942;
  Ncats(i1-i8) = 3;        // 3 valid responses
  Model(i1-i8) = GPC(3);

<Constraints>
  FIX (i1-i8), slope;           // do not estimate the slopes on GPC
  VALUE (i1-i8), slope, 1.0;    // constrain the slopes to equal 1
  FREE COV(1,1);              // estimate latent variable variance

```

is the corresponding parameter number. In the following section (Miscellaneous Control Values), we have a combination of pre- and post-execution information. The line Number of free parameters: specifies the number of parameters to be estimated for this calibration (two transition location parameters, δ_{j1} and δ_{j2} , times 8 items plus a common discrimination that is rescaled to be 1.0), followed by the number of iteration cycles that were executed (Number of cycles completed) and processing times. Because 21 cycles is less than the default 500, we know that we have a converged solution. flexMIRT v. 3.51 does not echo the first case (or two), nor does it indicate the number of lines read from the data file. Moreover, it does not provide any item descriptive statistics that would allow one to verify the data were correctly read.

In the Output Files section, we find the filenames corresponding to our request that the item parameter and person location estimates as well as the total information function (savePRM, saveSCO, saveINF) be saved. For example, the item parameter estimates and information function values are saved to files in which '-prm' and '-imf,' respectively, are appended to our command file's name (PCcalib) and the extension 'txt' is added. (Our person location estimates are saved to a '-sco' file (e.g., PCcalib-sco.txt), although this is not stated in the Output Files section.)

The Convergence and Numerical Stability section provides information

about the estimation's mathematical characteristics. The First-order test indicates whether the difference between iterations is less than the convergence criterion. Stated another way, there's no point in executing any more iterations because the difference between iterations is essentially "zero." Above we compared the number of executed iterations to the maximum allowed to determine that we had achieved convergence. The result of the First-order test (Convergence criteria satisfied) explicitly states that we have achieved convergence. As the name MMLE indicates, we wish to find the maximum of a function. However, as stated in Appendix A, a function may also have a minimum (or minima). The convergence criterion can be satisfied when the algorithm has found a maximum or a minimum. Therefore, the Second-order test determines whether we have a maximum or a minimum. As mentioned in Chapter 4 Endnote 10, when the Second-order test result is the Solution is a possible local maximum, we have a "sensible" solution.

Although our estimates are found in the subsequent table, we should go to the bottom of the output to examine the model-level fit information. The previous chapters' discussions of these statistics apply here and will not be repeated except to note that although the RMSEA of 0.05 indicates a "good fit," the TLI = 0.82 reflects a poor improvement in fit relative to the null model (assuming that we can generalize the Hu and Bentler [1999] results to a TLI based on M2). Returning to our item parameter estimates table, we see that each item occupies its own line, with the Label column containing the names provided with the Varnames command. As can be seen, our discrimination (*a*) is set to 1.0, with the subsequent columns containing the item location estimate $\hat{\delta}_j$ (labeled *b*), its standard error (s.e.), and offsets \hat{A}_h labeled *d* (e.g., *d* 2, *d* 3) with their corresponding standard errors. Given the parameterization used in flexMIRT, our $\hat{\delta}_{jh} = \hat{\delta}_j - \hat{A}_h$. That is, to obtain our $\hat{\delta}_{j1}$ and $\hat{\delta}_{j2}$ we combine item *j*'s $\hat{\delta}_j$ with its *d*2 and *d*3 values. For example, for item 1 $\hat{\delta}_{11} = \hat{\delta}_{j1} - d_2 = -1.679 - (-0.381) = -1.298$ and $\hat{\delta}_{12} = \hat{\delta}_{j1} - d_3 = -1.679 - 0.381 = -2.060$. For items 2–8 we have $\hat{\delta}_{21} = -0.428$ and $\hat{\delta}_{22} = -1.331$, $\hat{\delta}_{31} = -1.041$ and $\hat{\delta}_{32} = -0.233$, $\hat{\delta}_{41} = -0.279$ and $\hat{\delta}_{42} = 0.599$, $\hat{\delta}_{51} = 1.587$ and $\hat{\delta}_{52} = -1.218$, $\hat{\delta}_{61} = -0.181$ and $\hat{\delta}_{62} = 2.609$, $\hat{\delta}_{71} = 1.600$ and $\hat{\delta}_{72} = -0.011$, and $\hat{\delta}_{81} = 1.269$ and $\hat{\delta}_{82} = 1.366$.

As mentioned above, the GPC model is implemented as a special case of the nominal model. The tables following our estimates contain the transformation matrices' information associated with transforming and implementing constraints on the nominal model's estimates. (In the tables, the [category] discrimination/slopes are labeled either *a* or alpha, and the intercepts are labeled either *c* or gamma depending on the table.) In the Nominal Model Slopes and Scoring Function Contrasts table, we see that the overall slope (*a*) has the specified value of 1.000, the transformation matrix for our slopes (*Ta*) is using trend contrasts, and the $3 - 1 = 2$ (i.e., $m_j - 1$) contrast estimates (alpha 1, alpha 2) and their corresponding standard errors for each of our items. Our Nominal Model Scoring Function Values table contains information that requires that we briefly mention a generalization of Bock's nominal response model. Specifically, Thissen, Cai, and Bock (2010) re-parameterized Bock's nominal model to have a single (overall) discrimination parameter in addition to category discrimina-

TABLE 7.2. Abridged Output from the flexMIRT Rasch PC Model Calibration of the Alike Data

flexMIRT(R) Engine Version 3.51 (64-bit)
Flexible Multilevel Multidimensional Item Response Modeling and Test Scoring
(C) 2013-2017 Vector Psychometric Group, LLC., Chapel Hill, NC, USA

PC calibration, ALIKE data - slope = 1
Item & person parameter estimates - single run

Summary of the Data and Dimensions

Missing data code	-9
Number of Items	8
Number of Cases	2942
# Latent Dimensions	1

Item	Categories	Model	Ta	Tc
1	3	Nominal	Trend	Trend
2	3	Nominal	Trend	Trend
:				
8	3	Nominal	Trend	Trend

Bock-Aitkin EM Algorithm Control Values

Maximum number of cycles: 500
Convergence criterion: 1.00e-004
Maximum number of M-step iterations: 100
Convergence criterion for iterative M-steps: 1.00e-007
Number of rectangular quadrature points: 49
Minimum, Maximum quadrature points: -6.00, 6.00
Standard error computation algorithm: Fisher (Expected)

User-Defined Initial Values for Parameters

P#	Value
18	1.00
21	1.00
24	1.00
27	1.00
30	1.00
33	1.00
36	1.00
39	1.00

Miscellaneous Control Values

Z tolerance, max. abs. logit value: 50.00
Number of free parameters: 17
Number of cycles completed: 21
Number of processor cores used: 1
Maximum parameter change (P#): 0.000097778 (17)

Processing times (in seconds)

E-step computations: 0.09
M-step computations: 0.01
Standard error computations: 0.72
Goodness-of-fit statistics: 0.03
Total: 0.86

(continued)

TABLE 7.2. (continued)

Output Files

Text results and control parameters: PCcalib-irt.txt

Text parameter estimate file: PCcalib-prm.txt

Information values in a file: PCcalib-inf.txt

Convergence and Numerical Stability

flexMIRT(R) engine status: Normal termination

First-order test: Convergence criteria satisfied

Condition number of information matrix: 8.8708

Second-order test: Solution is a possible local maximum

*** Random effects calibration in Group 1: OnlyGroup

:

GPC Items for Group 1: OnlyGroup

Item	Label	P#	a	s.e.	b	s.e.	d	1	d	2	s.e.	d	3	s.e.
1	i1		1.000		-1.679	0.046	0	-0.381	0.066	0.381	0.066			
2	i2		1.000		-0.880	0.033	0	-0.452	0.052	0.452	0.052			
3	i3		1.000		-0.637	0.033	0	0.404	0.042	-0.404	0.042			
4	i4		1.000		0.160	0.031	0	0.439	0.039	-0.439	0.039			
5	i5		1.000		0.185	0.028	0	-1.402	0.066	1.402	0.066			
6	i6		1.000		1.214	0.046	0	1.395	0.049	-1.395	0.049			
7	i7		1.000		0.794	0.032	0	-0.805	0.057	0.805	0.057			
8	i8		1.000		1.317	0.040	0	0.049	0.052	-0.049	0.052			

:

Nominal Model Slopes & Scoring Function Contrasts for Items for Group 1: OnlyGroup

Item	Label	P#	a	s.e.	Contrasts	P#	alpha	1	s.e.	P#	alpha	2	s.e.
1	i1		1.000	----	Trend		1.000	----	0.000	0.000	----		
2	i2		1.000	----	Trend		1.000	----	0.000	0.000	----		
:													
8	i8		1.000	----	Trend		1.000	----	0.000	0.000	----		

Nominal Model Scoring Function Values Group 1: OnlyGroup

Item	Label	s	1	s	2	s	3
1	i1		0.000		1.000		2.000
2	i2		0.000		1.000		2.000
:							
8	i8		0.000		1.000		2.000

Nominal Model Intercept Contrasts for Items for Group 1: OnlyGroup

Item	Label	Contrasts	P#	gamma	1	s.e.	P#	gamma	2	s.e.
1	i1	Trend	1	1.679	0.046		2	-0.381	0.066	
2	i2	Trend	3	0.880	0.033		4	-0.452	0.052	
:										
8	i8	Trend	15	-1.317	0.040		16	0.049	0.052	

Original (Bock, 1972) Parameters, Nominal Items for Group 1: OnlyGroup

Item	Label	Category:	1	2	3
1	i1	a	0.000	1.000	2.000
		c	0.000	1.298	3.359
2	i2	a	0.000	1.000	2.000
		c	0.000	0.428	1.759
:					

(continued)

TABLE 7.2. (continued)

8	i8	a	0.000	1.000	2.000					
		c	0.000	-1.269	-2.635					
:										
Group Parameter Estimates:										
:										
Marginal fit (Chi-square) and Standardized LD X2 Statistics for Group 1: OnlyGroup										
Marginal										
Item	Chi2	1	2	3	4	5	6	7		
1	0.1									
2	0.1	64.5p								
3	0.1	45.9p	59.2n							
4	0.1	7.6n	3.9n	16.5p						
5	0.2	9.7p	17.2p	36.2n	10.8n					
6	0.1	3.7n	4.8n	19.5n	13.6n	1.7n				
7	0.0	3.6p	6.0p	17.8p	1.1p	16.3n	9.7n			
8	0.1	1.9n	2.9n	8.5n	9.6n	13.8n	9.6p	12.3n		
:										
Item Information Function Values at 15 Values of theta from -2.8 to 2.8 for Group 1: OnlyGroup										
Theta:										
Item	Label	-2.8	-2.4	-2.0	-1.6	...	1.6	2.0	2.4	2.8
1	i1	0.38	0.55	0.70	0.74	...	0.03	0.02	0.01	0.01
2	i2	0.15	0.24	0.38	0.56	...	0.07	0.04	0.03	0.02
:										
8	i8	0.02	0.03	0.04	0.06	...	0.63	0.53	0.40	0.27
Test Information: 2.31 2.74 3.27 3.81 ... 3.71 3.14 2.67 2.33										
Expected s.e.: 0.66 0.60 0.55 0.51 ... 0.52 0.56 0.61 0.66										
Marginal reliability for response pattern scores: 0.70										
Statistics based on the loglikelihood of the fitted model:										
-2loglikelihood: 39370.52										
Akaike Information Criterion (AIC): 39404.52										
Bayesian Information Criterion (BIC): 39506.30										
Statistics based on the loglikelihood of the zero-factor null model:										
-2loglikelihood: 41755.85										
Akaike Information Criterion (AIC): 41787.85										
Bayesian Information Criterion (BIC): 41883.64										
Full-information fit statistics of the fitted model:										
Degrees										
G2 of freedom Probability F0hat RMSEA										
3148.60 886 0.0001 1.0702 0.03										
The table is too sparse to compute the Pearson X2 statistic.										
Even though G2 is shown, it should be interpreted with caution.										
Full-information fit statistics of the zero-factor null model:										
Degrees										
G2 of freedom Probability F0hat RMSEA										
5533.93 887 0.0001 1.8810 0.04										

(continued)

TABLE 7.2. (continued)

The table is too sparse to compute the Pearson X² statistic.
Even though G₂ is shown, it should be interpreted with caution.

With the zero-factor null model, Tucker-Lewis (non-normed) fit index based on G₂ is 0.51

Limited-information fit statistics of the fitted model:

	Degrees	M2 of freedom	Probability	F0hat	RMSEA
	1064.30	111	0.0001	0.3618	0.05

Note: M₂ is based on full marginal tables.

Note: Model-based weight matrix is used.

Limited-information fit statistics of the zero-factor null model:

	Degrees	M2 of freedom	Probability	F0hat	RMSEA
	5332.09	112	0.0001	1.8124	0.13

Note: M₂ is based on full marginal tables.

Note: Model-based weight matrix is used.

With the zero-factor null model, Tucker-Lewis (non-normed) fit index based on M₂ is 0.82

The following are the EAP $\hat{\theta}$ s found in the saved 'sco.txt' file (i.e., PCcalib-sco.txt). The format is Group, record's ordinal position in the data file, $\hat{\theta}$, and PSD($\hat{\theta}$).

1	1	-0.791483	0.465536
1	2	-0.791483	0.465536
1	3	0.025355	0.443908
1	4	0.420576	0.446935
1	5	0.222305	0.444089
1	6	1.539544	0.515857
:			
1	2939	-0.578945	0.456880
1	2940	0.420576	0.446935
1	2941	1.285397	0.493176
1	2942	1.051198	0.475437

tions and introduced the use of $m - 1$ scoring functions to relate the value/order of the categories to our θ ; also see Thissen and Cai (2016). In our unidimensional case, we have a scoring function vector whose values are provided in this table. (Thissen et al.'s reparameterization generalizes Bock's model to be multidimensional in which we have a θ vector and a matrix of scoring functions.) The Nominal Model Intercept Contrasts table contains analogous information to that found in the Nominal Model Slopes . . . table, but for the Tc transformation matrix used with the intercepts (i.e., contrast type, the intercept (gamma) estimates, and their corresponding standard errors). These are the estimates that are saved to the item parameter estimate file (i.e., '-prm.txt') and that are transformed to produce the estimates we discussed above. The last estimate table, Original (Bock, 1972) Parameters, contains the corresponding nominal response model estimate in which each category has slope (a) and intercept (c) parameter estimates.

To examine local independence, we use Chen and Thissen's (1997) χ^2 LD; additional item-level fit information, such as Orlando and Thissen's S – χ^2 (2000, 2003), can be obtained by setting GOF to Complete. χ^2 LD is calculated for each item pair's two-way contingency table. However, the values presented are not χ^2 s. With polytomous data, it is possible to have item pairs with a different number of response categories. For example, one item pair might have five response categories and another might have three. As such, flexMIRT "standardizes" the values to facilitate comparing across item pairs. Moreover, with our ordered responses, flexMIRT calculates the correlation for each item pair based on the observed data and the model. If the observed data correlation is greater than the model expected value, then the suffix 'p' is used; otherwise the suffix 'n' is attached. Our standardized values are found in the Marginal fit (Chi-square) and . . . table; these values are approximately z-scores (Cai, 2013). As was the case with Q_3 and Q_3^P , we are looking for "large" values.⁹ We see, for instance, that for the item pair 1–2 the model-based expectation is substantially less than what was observed, and the converse is true for the item pair 2–3. Item pairs 1–3 and 3–5 also exhibit large values. As was done in the previous chapter, we view these indices as indicative of potential problems that should be further investigated rather than from a null hypothesis significance testing perspective. At a minimum, these four item pairs should be examined as discussed in Chapter 6; it would be prudent to also examine all item pairs whose values exceed 10 (e.g., item pairs 2–5, 3–6, and 3–7). Above we described what this examination would entail. However, specific to this statistic, our investigation would also seek to determine whether the corresponding m x m tables are fully populated or whether it is the case that we have several zero frequency cells that would inflate the χ^2 .

The following section of the output contains information function information. The first section contains each item's information. The Theta: line provides the point on the continuum at which the subsequent lines' information value occurs. For example, at -2.8 item 1's (i1) information is 0.38 and item 2's (i2) is 0.15; at -2.4 item 1's information is 0.55 and item 2's is 0.24, and so on. Following this section is the instrument's total information function (Test Information:) and the corresponding standard error (Expected s.e.); the Theta: scale is the same as the item information.

As discussed in Chapter 5, sometimes it is desirable to have a single bounded value that represents the quality of estimation for the entire continuum. The marginal reliability is an index with a range from 0 to 1, with values that approach or are equal to 1.0, reflecting small error variability. From our $\hat{\theta}$ s found in the '-sco.txt' file we calculate the average error variance to be 0.210622 and our $s_e^2 = 0.491181$. From Equation 5.25 we calculate our estimated marginal reliability to be

$$\hat{\rho} = \frac{s_{\hat{\theta}}^2}{s_{\hat{\theta}}^2 + \text{avg}(s_e^2(\hat{\theta}))} = \frac{0.491181}{0.491181 + 0.210622} = 0.69988$$

flexMIRT reports the marginal reliability as 0.70. Although a single accuracy value may be desirable, the trade-off is that it may potentially be misinterpreted. For instance, given the non-uniformity in the total information function (Figure 7.6), this empirical

reliability understates the accuracy in the center of the metric and overstates the estimation accuracy for (approximately) $\hat{\theta}_s < -2$ and $\hat{\theta}_s > 2$. Thus, it is only when the total information function is somewhat uniformly distributed that this value accurately characterizes the precision of measurement across the continuum.

flexMIRT does not directly provide graphics but does provide an R script `flexmirt.R`. This file contains several R functions to be used to create IRFs/ORFs, TCCs, item and total information function plots. (At present it is not possible to select the plots for just a single item without modifying the provided code.) For our example we first specify the folder that contains our saved `PCcalib-prm.txt` and `PCcalib-inf.txt` files by using the R `setwd("C:\<subdirectory name>")` function. Subsequently, we specify the path to where the `flexmirt.R` file is located (e.g., `source("C:\\flexMIRT\\flexmirt.R")`).

To obtain our ORFs and TCC plots, we call the `flexmirt.icc` function (`flexmirt.icc("PCcalib-prm.txt", theta = seq(-4,4,0.1), savepdf = F)`) specifying a range of -4 to 4 in 0.1 increments and to not save the plots to a pdf file. Figure 7.4 contains the ORFs and information functions for item 1. Given the reversal in $\hat{\delta}_{1h}$ for item 1 (i.e., $\hat{\delta}_{11} = -1.298$ and $\hat{\delta}_{12} = -2.060$), it is not surprising that respondents located above (approximately) -2.0 are more likely to receive full credit (ORF labeled "2") rather than partial credit (ORF labeled "1"; left panel of Figure 7.4). Similarly, respondents located below -2.0 have a higher probability of receiving no credit (ORF labeled "0") rather than partial credit. The ORFs show that, in effect, this item is primarily behaving like a dichotomous item. For most of the respondents, this is a relatively easy item. This item's information function is obtained by calling the function `flexmirt.inf("PCcalib-inf.txt", savepdf = F)`. The right panel of Figure 7.4 shows that item 1 provides its maximum information at approximately -1.7.

Figure 7.5 contains the ORFs for all the items on the instrument. Four of the items have a tendency to behave in a dichotomous way (items 1, 2, 5, and 7), whereas items 3,

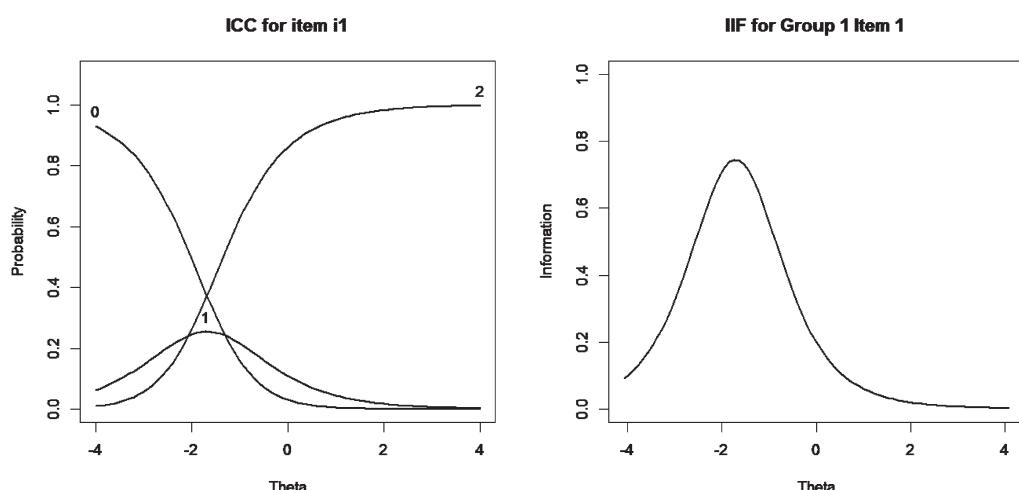


FIGURE 7.4. ORFs and item information function for item 1.

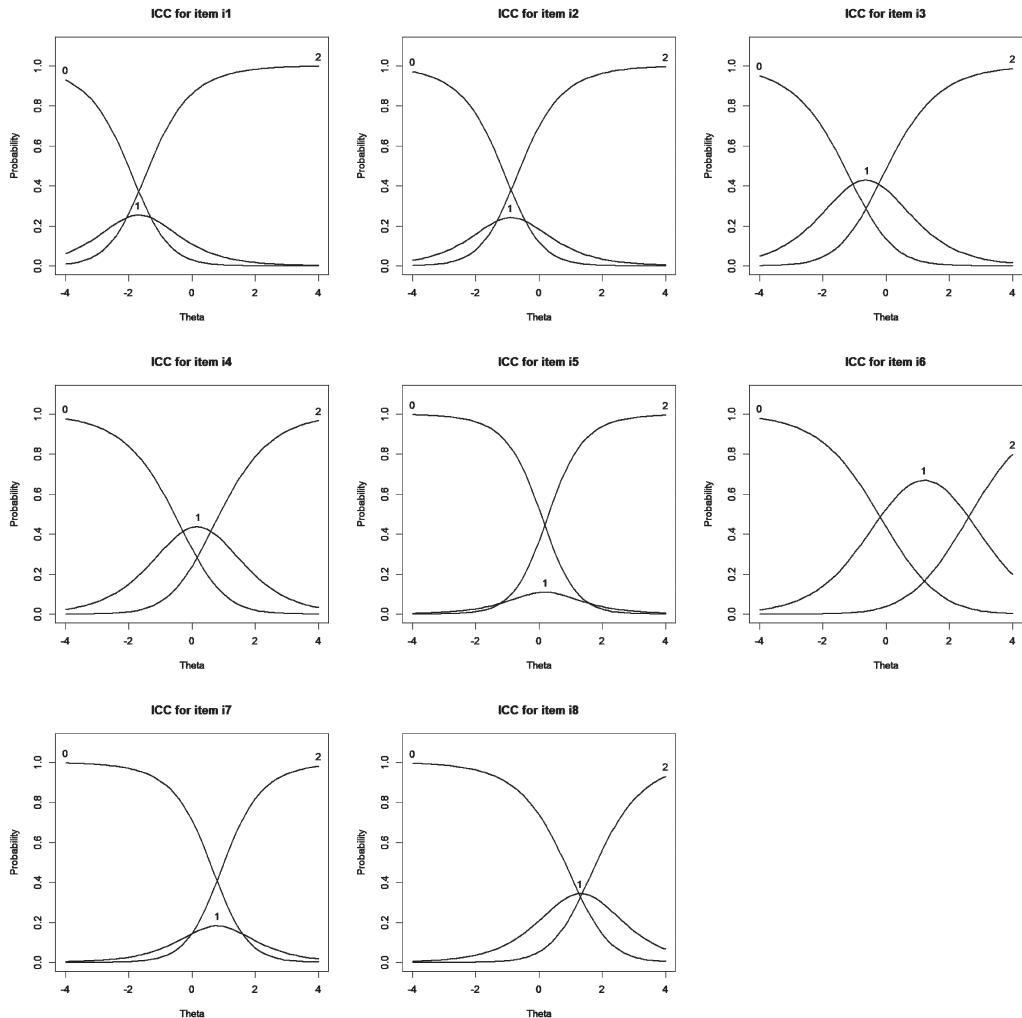


FIGURE 7.5. ORFs for all eight Alike items.

4, 6, and 8 are trichotomous. It is also seen that items 1, 2, and 3 are useful for assessing people located in the lower half of the θ continuum, items 4, 5, and 7 provide information in the middle of the continuum, and items 6 and 8 are tapping the upper half of the continuum. Therefore, the instrument gives relatively broad capacity to measure reasoning ability.

The sum of the item information functions (e.g., right panel Figure 7.4) is captured by the instrument's total information (in Table 7.2, labeled Test Information). A graphical depiction of this information function is shown in Figure 7.6. As can be seen, the accuracy of person location estimation varies as a function of the specific location. Specifically, this instrument provides comparatively more accurate estimates of a person's location from around -0.5 to 1.0 and a progressively less accurate estimate as one approaches 2 or -2 and beyond. (The abscissa scale is determined from the values pro-

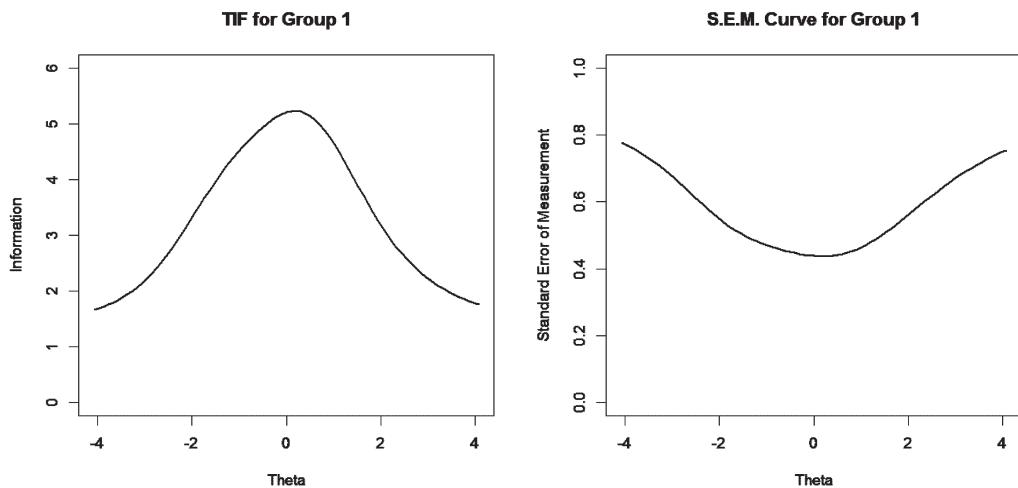


FIGURE 7.6. Total information for Alike reasoning exam, PC model (left) and expected standard error (right).

vided in the `FisherInf` command; we specified a range of -4 to 4 . This implies that because flexMIRT reads the command file to obtain this information, the command file needs to be in the same folder as the '`-inf.txt`' file.)

Example: Application of the PC Model to a Reasoning Ability Instrument, MMLE, mirt

Although several R packages (e.g., `eRm`, `mixRasch`, `mirt`) can be used to perform a PC model calibration of our data, we use `mirt` for our PC model calibration; `eRm` is used in Appendix E, and `mixRasch` is used with the Rating Scale model below. As in the previous chapter, we assume the relevant libraries are loaded into our R workspace; Table 7.3 shows our session. To verify that our data are correctly read, we use the `head` and `tail` functions. As can be seen, all responses codes fall between 0 and 2 (inclusive) obtained by using the `describe(alikedata)` function from the `Hmisc` package. In addition, because we have polytomous data, we should determine the extent to which all response categories were used because categories with very small or zero frequencies will make it difficult or impossible to estimate the corresponding item parameters. As can be seen, we do not have any item with null categories and mostly “largish” frequencies for our response categories across items.

Although we do not repeat the analyses that we have performed in previous chapters (e.g., dimensionality, invariance), it should be understood that those analyses are routinely conducted and have been done for this calibration. To perform our Rasch PC calibration, we use the `Rasch` `itemtype` argument (`itemtype = 'Rasch'`).¹⁰ By specifying `itemtype` as `Rasch`, we also obtain the traditional item fit statistics. Our calibration required 21 iterations to obtain convergence. Our above discussion of the

TABLE 7.3. mirt Session for the Rasch PC Model Calibration of the Alike Data

```

> library(Hmisc)
> library(psych)
> library(ggplot2)
> library(mirt)
> alikedata = read.table(file.choose(), header=FALSE)

> names(alikedata) = c(paste0("I",1:8))
> head(alikedata,5)
  I1 I2 I3 I4 I5 I6 I7 I8
1  1  2  0  0  0  0  0  0
2  2  2  0  0  0  0  0  0
3  2  0  2  1  2  1  0  0
4  2  2  1  2  2  1  0  0
5  2  2  2  1  2  0  0  0

> tail(alikedata,5)
  I1 I2 I3 I4 I5 I6 I7 I8
2938 1  0  2  1  0  1  0  1
2939 2  2  0  0  0  1  0  0
2940 2  2  1  1  2  1  1  0
2941 2  2  2  2  2  2  2  0
2942 2  1  1  2  2  2  2  1

> Hmisc::describe(alikedata)
alikedata

 8 Variables      2942 Observations
-----
I1
  n    missing  distinct      Info      Mean      Gmd
  2942        0         3     0.462     1.744     0.4323

Value      0      1      2
Frequency   201   351  2390
Proportion 0.068 0.119 0.812
-----
I2
  n    missing  distinct      Info      Mean      Gmd
  2942        0         3     0.715     1.468     0.7533

Value      0      1      2
Frequency   536   494  1912
Proportion 0.182 0.168 0.650
-----
I3
  n    missing  distinct      Info      Mean      Gmd
  2942        0         3     0.845     1.297     0.7978

Value      0      1      2
Frequency   537   995 1410
Proportion 0.183 0.338 0.479
-----
I4
  n    missing  distinct      Info      Mean      Gmd
  2942        0         3     0.884     0.9283    0.8561

Value      0      1      2
Frequency  1028  1097  817
Proportion 0.349 0.373 0.278

```

(continued)

TABLE 7.3. (continued)

I5					
n missing distinct		Info	Mean	Gmd	
2942 0 3		0.802	0.8923	0.981	
Value 0 1 2					
Frequency 1500 259 1183					
Proportion 0.510 0.088 0.402					
I6					
n missing distinct		Info	Mean	Gmd	
2942 0 3		0.79	0.6186	0.6065	
Value 0 1 2					
Frequency 1299 1466 177					
Proportion 0.442 0.498 0.060					
I7					
n missing distinct		Info	Mean	Gmd	
2942 0 3		0.707	0.5588	0.7884	
Value 0 1 2					
Frequency 1927 386 629					
Proportion 0.655 0.131 0.214					
I8					
n missing distinct		Info	Mean	Gmd	
2942 0 3		0.655	0.4018	0.5985	
Value 0 1 2					
Frequency 2043 616 283					
Proportion 0.694 0.209 0.096					

```
> print((raschpcm = mirt(alikedata,model=1,itemtype='Rasch',SE=T,SE.type='Fisher')))
  Iteration: 21, Log-Lik: -19685.262, Max-Change: 0.00008

  Calculating information matrix...

  Call:
  mirt(data = alikedata, model = 1, itemtype = "Rasch", SE = T,
    SE.type = "Fisher")

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 21 EM iterations.
  mirt version: 1.30
  M-step optimizer: nlminb
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Information matrix estimated with method: Fisher
  Condition number of information matrix = 50195.05
  Second-order test: model is a possible local maximum

  Log-likelihood = -19685.26
  Estimated parameters: 17
  AIC = 39404.52; AICc = 39404.73
  BIC = 39506.3; SABIC = 39452.28
  G2 (6543) = 3148.6, p = 1
  RMSEA = 0, CFI = NaN, TLI = NaN
```

(continued)

TABLE 7.3. (continued)

```

> M2(raschpcm,CI=0.95)
      M2 df p      RMSEA   RMSEA_2.5
stats 300.8585 19 0 0.07102173 0.06270174
      RMSEA_97.5    SRMSR      TLI      CFI
stats 0.07954564 0.0632744 0.8400804 0.8480763

> residuals(raschpcm)
LD matrix (lower triangle) and standardized values:

      I1      I2      I3      I4      I5      I6      I7      I8
I1     NA  0.178  0.151 -0.066  0.073 -0.050  0.049 -0.040
I2  186.404  NA -0.171 -0.051  0.095 -0.055  0.060 -0.045
I3 133.753 -171.383  NA  0.093 -0.134 -0.100  0.096 -0.069
I4 -25.527 -15.037  50.714  NA -0.077 -0.085  0.035 -0.073
I5  31.462  52.597 -106.437 -34.604  NA -0.039  0.092 -0.086
I6 -14.484 -17.591 -59.252 -42.631 -8.865  NA -0.073  0.073
I7  14.257  21.062  54.242  7.121  50.203 -31.436  NA  0.081
I8 -9.496 -12.121 -27.924 -31.127 -43.122  31.255  38.712  NA
>

> itemfit(raschpcm,group.bins=10,empirical.plot=1)                      # produces Figure 7.7A
> itemfit(raschpcm,group.bins=6,empirical.plot=1)                         # produces Figure 7.7B

> coef(raschpcm,simplify=TRUE,IRTpars=TRUE)
$items
  a      b1      b2
I1 1 -1.299 -2.060
I2 1 -0.428 -0.428
I3 1 -1.041 -0.233
I4 1 -0.279  0.599
I5 1  1.587 -1.217
I6 1 -0.180  2.610
I7 1  1.600 -0.010
I8 1  1.269  1.366

$means
Theta
  0

$cov
  Theta
Theta  1

> # obtain person estimates & display first 6 cases
>
head((people_raschpcm=fscores(raschpcm,method="EAP",full.scores=T,full.scores.SE=T),6)
      F1      SE_F1
[1,] -0.79166238 0.4656082
[2,] -0.79166238 0.4656082
[3,]  0.02540245 0.4439650
[4,]  0.42072392 0.4469928
[5,]  0.22240226 0.4441454
[6,]  1.54005352 0.5159673

> tail(people_raschpcm,4)
      F1      SE_F1
[2939,] -0.5790602 0.4569465
[2940,]  0.4207239 0.4469928
[2941,]  1.2858056 0.4932671
[2942,]  1.0515256 0.4755144

```

(continued)

TABLE 7.3. (*continued*)

```

> mean(people_raschpcm[,1])                                     # average person est
[1] 3.273777e-06
> sd(people_raschpcm[,1]))                                    # SD person est
[1] 0.7010452

> # save person estimates to external csv formatted file
> write.csv(people_raschpcm, file = "peopleRaschPC_EAP.csv")

> marginal_rxx(raschpcm)
[1] 0.7552859

> # empirical reliability
> fscores(raschpcm,method="EAP",full.scores=T,full.scores.SE=T,returnER=T)
      F1
0.6999438

># person fit information
> head((people_raschpcmFit=personfit(raschpcm,method="EAP")) , 6)
      outfit    z.outfit     infit    z.infit      Zh
1 0.5501197 -0.5627299 0.6283094 -0.6782813 0.05076339
2 0.5779249 -0.5032595 0.7411920 -0.3918528 1.24150469
3 1.1221237  0.3962376 1.2477131  0.6731525 0.29375536
4 0.5591348 -0.5969187 0.6703643 -0.7689996 1.11334487
5 0.5489978 -0.7108036 0.6008066 -0.9836717 1.31072607
6 0.8990370  0.3081714 1.0843042  0.3412464 -0.76520223

> tail(people_raschpcmFit,4)
      outfit    z.outfit     infit    z.infit      Zh
2939 0.6979446 -0.34597494 0.7900259 -0.28867323 0.9909814
2940 0.3291154 -1.21294955 0.3726715 -1.89908132 0.5233724
2941 0.7407936  0.06932336 0.9397832  0.07509505 0.5480734
2942 1.5941598  0.85776959 1.1199106  0.39821177 -1.6101236

> psych::describe(people_raschpcmFit)                         # to obtain skew & kurtosis
      vars     n   mean    sd median trimmed  mad   min   max range skew
outfit      1 2942  0.80  0.58    0.66    0.70  0.35  0.17  9.01  8.83  3.61
z.outfit     2 2942 -0.22  0.80   -0.35   -0.30  0.69 -1.80  3.87  5.67  1.16
infit       3 2942  0.85  0.44    0.74    0.79  0.37  0.18  3.69  3.51  1.27
z.infit      4 2942 -0.35  1.02   -0.45   -0.39  1.00 -2.90  3.71  6.61  0.36
Zh          5 2942  0.28  0.96    0.44    0.39  0.91 -4.55 1.60  6.16 -1.12

      kurtosis    se
outfit      27.13 0.01
z.outfit     2.18 0.01
infit       2.22 0.01
z.infit     -0.01 0.02
Zh          1.43 0.02

> # dot plot - Figure 7.8 (ggplot2 library)
> ggplot(people_raschpcmFit,aes(x=infit)) + geom_dotplot(binwidth=.05)
> ggplot(people_raschpcmFit,aes(x=outfit))+ geom_dotplot(binwidth=.05)

> alikedataSort= alikedata[order(avgdEst)]                  # sort items by mean item loc

> alikedataSort$X=rowSums(alikedataSort[,1:8] )            # add observed score X

> alikedataSort$Id=seq(1,2942,1)                            # add case id
> alikedataSort$outfit=people_raschpcmFit$outfit           # add outfit stat
> alikedataSort$z.outfit=people_raschpcmFit$z.outfit         # add z.outfit stat
> alikedataSort$infit=people_raschpcmFit$infit             # add infit stat
> alikedataSort$z.infit=people_raschpcmFit$z.infit           # add z.infit stat

```

(continued)

TABLE 7.3. (continued)

```

> # the next two steps would not be necessary if we had initially set full.scores=F
> # our call to fscore
> alikedataSort$t_est=people_raschpcm[,1]                                # add (merge) EAP est
> alikedataSort$se=people_raschpcm[,2]                                     # add (merge) std err

> # sort by infit
> head(pFitsortINFIT=alikedataSort[order(alikedataSort$infit),],6)
    I1 I2 I3 I4 I5 I7 I6 I8 X     id   outfit z.outfit      infit
  510  2  2  2  1  2   1   1   1 12   510 0.1727697 -1.3805 0.1775261
  663  2  2  2  1  2   1   1   1 12   663 0.1727697 -1.3805 0.1775261
  918  2  2  2  1  2   1   1   1 12   918 0.1727697 -1.3805 0.1775261
  960  2  2  2  1  2   1   1   1 12   960 0.1727697 -1.3805 0.1775261
  969  2  2  2  1  2   1   1   1 12   969 0.1727697 -1.3805 0.1775261
 1464  2  2  2  1  2   1   1   1 12  1464 0.1727697 -1.3805 0.1775261
          z.infit      t_est      se
  510 -2.763097 0.8320015 0.4622223
  663 -2.763097 0.8320015 0.4622223
  918 -2.763097 0.8320015 0.4622223
  960 -2.763097 0.8320015 0.4622223
  969 -2.763097 0.8320015 0.4622223
 1464 -2.763097 0.8320015 0.4622223

> tail(pFitsortINFIT,4)
    I1 I2 I3 I4 I5 I7 I6 I8 X     id   outfit z.outfit      infit z.infit
  664  0  0  2  2  0   2   1   1 8   664 4.084325 3.366506 2.963367 3.155486
 1800  0  0  0  2  2   1   1   0 6   1800 2.984636 2.497645 3.015394 2.927535
  329  0  0  2  2  2   2   0   0 8   329 4.115695 3.388184 3.075811 3.279275
  244  0  0  0  2  2   2   0   1 7   244 4.216501 3.509058 3.688478 3.714341
          t_est      se
  664  0.02540245 0.4439650
 1800 -0.37335088 0.4505023
  329  0.02540245 0.4439650
  244 -0.17250870 0.4461523

> # sort by outfit
> head(pFitsortOUTFIT=alikedataSort[order(alikedataSort$outfit),],6)
    I1 I2 I3 I4 I5 I7 I6 I8 X     id   outfit z.outfit      infit
  510  2  2  2  1  2   1   1   1 12   510 0.1727697 -1.3805 0.1775261
  663  2  2  2  1  2   1   1   1 12   663 0.1727697 -1.3805 0.1775261
  918  2  2  2  1  2   1   1   1 12   918 0.1727697 -1.3805 0.1775261
  960  2  2  2  1  2   1   1   1 12   960 0.1727697 -1.3805 0.1775261
  969  2  2  2  1  2   1   1   1 12   969 0.1727697 -1.3805 0.1775261
 1464  2  2  2  1  2   1   1   1 12  1464 0.1727697 -1.3805 0.1775261
          z.infit      t_est      se
  510 -2.763097 0.8320015 0.4622223
  663 -2.763097 0.8320015 0.4622223
  918 -2.763097 0.8320015 0.4622223
  960 -2.763097 0.8320015 0.4622223
  969 -2.763097 0.8320015 0.4622223
 1464 -2.763097 0.8320015 0.4622223

> pFitsortzINFIT=alikedataSort[order(alikedataSort$z.infit),]           # sort-z.infit
> pFitsortzOUTFIT=alikedataSort[order(alikedataSort$z.outfit),]           # sort-z.outfit

> # extract item param ests
> dEst=coef(raschpcm,simplify=TRUE,IRTpars=TRUE)$'items'[, -1]
> dEst
      b1          b2
  I1 -1.2988057 -2.06030224
  I2 -0.4280372 -1.33126847
  I3 -1.0412969 -0.23277796
  I4 -0.2785514  0.59887154

```

(continued)

TABLE 7.3. (continued)

```

I5  1.5867891 -1.21746490
I6 -0.1804882  2.60974592
I7  1.5997857 -0.01044884
I8  1.2689128  1.36625415

> avgdEst=rowMeans(dEst)                                     # calc average item location
> avgdEst
    I1          I2          I3          I4
-1.6795540 -0.8796528 -0.6370374  0.1601601
    I5          I6          I7          I8
0.1846621  1.2146288  0.7946684  1.3175835

> # item 6 & 7 out of order
> avgdEstSort=avgdEst[order(avgdEst)]                         # sort average item locations
> avgdEstSort
    I1          I2          I3          I4
-1.6795540 -0.8796528 -0.6370374  0.1601601
    I5          I7          I6          I8
0.1846621  0.7946684  1.2146288  1.3175835

> alikedataSort= alikedata[order(avgdEst)]                   # sort resp by avh item loc
> head(alikedataSort,6)
  I1 I2 I3 I4 I5 I7 I6 I8
1  1  1  2  0  0  0  0  0
2  2  2  0  0  0  0  0  0
3  2  0  2  1  2  0  1  0
4  2  2  1  2  2  0  1  0
5  2  2  2  1  2  0  0  0
6  2  2  2  2  1  2  2  2

> # sorted by X & OUTFIT
> pFitsortOUTFITX=alikedataSort[order(alikedataSort$X,alikedataSort$outfit),]

> # sorted by X & INFIT
> pFitsortINFITX=alikedataSort[order(alikedataSort$X,alikedataSort$outfit),]

```

interpretation of the information criteria and the M2 statistic for model-level fit analysis applies here and is not repeated.

As done in previous chapters (e.g., Chapter 4), we would obtain statistical and graphical item-level fit information. Suffice it to say that we would examine fit by examining the items' INFIT and OUTFIT values as done in Chapter 4 and would compare an item's empirical ORF with its predicted. Figure 7.7 shows one such example plot for polytomous data. As can be seen, regardless of whether we use 10 or 6 "bins" (e.g., `itemfit(. . .)`) we see agreement between the expected and observed data for item 1. In short, we do not repeat the discussion of the steps for item-level analysis except for the introduction of the use of χ^2 LD with mirt.

To parallel our flexMIRT calibration's use of χ^2 LD, we now use the `residuals` function to obtain conditional dependence information based on χ^2 LD (i.e., `residuals(raschpcm)`). The lower off-diagonal contains the χ^2 LDs, with the upper off-diagonal showing the corresponding standardized values ("Cramér's V"). Cramér's V can be interpreted as an "effect size" measure, with a value of 1 reflecting two perfectly

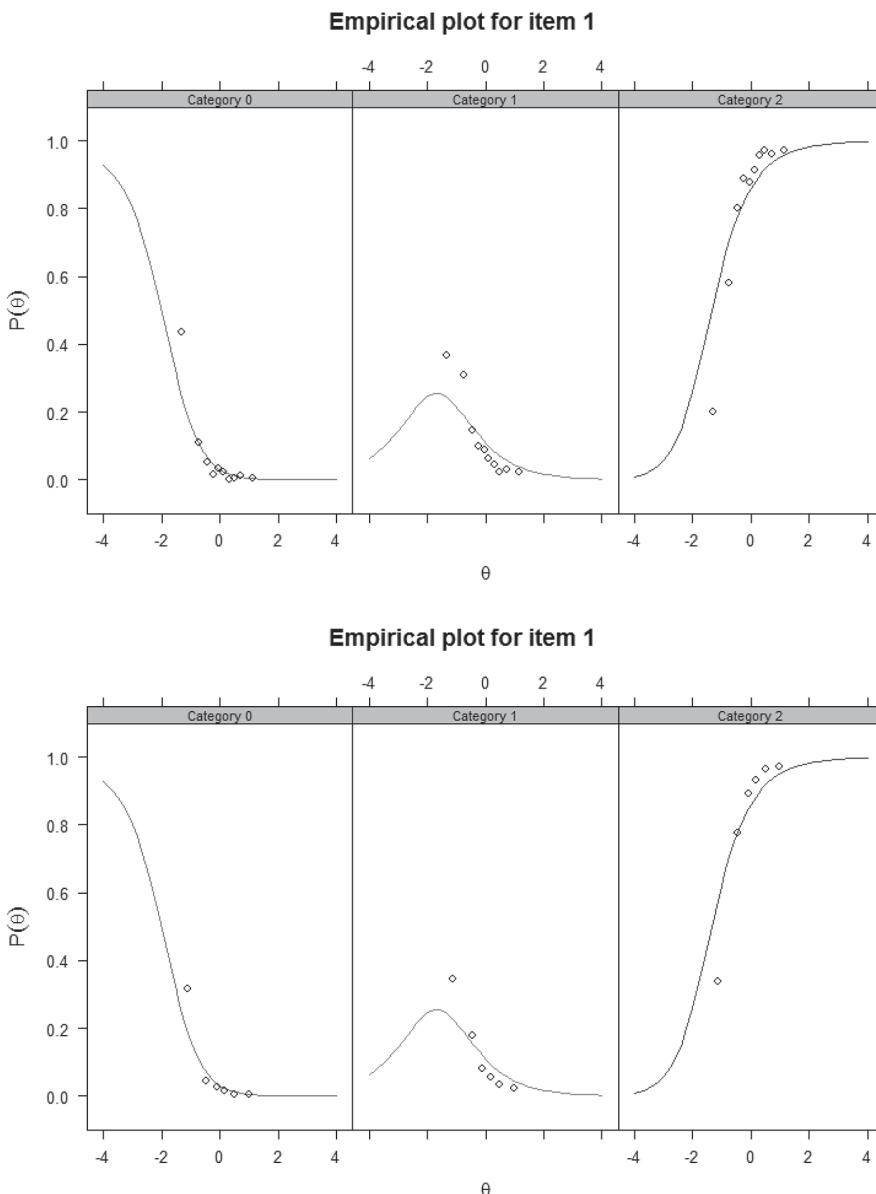


FIGURE 7.7. Empirical and predicted ORFs for item 1. Panel A (top): 10 bins, Panel B (middle): 6 bins.

dependent items and a $V = 0$ indicating independence of the items. Thus, values close to 1 indicate a strong relationship, with values close to 0 indicating negligible association between the two items.¹¹ As Cohen (1988) states, the best approach for interpreting effect sizes is the development of a context-specific sense of what should be considered to be a small, average, or large effect size. However, in the current context, such information does not exist. Therefore, we use Cohen's guidelines of what constitutes a small,

TABLE 7.4. Effect Sizes for Cramér's V

Label	m_j				
	2	3	4	5	6
small	0.100	0.071	0.058	0.05	0.045
medium	0.300	0.212	0.173	0.15	0.134
large	0.500	0.354	0.289	0.25	0.224

Note. If items vary in terms of the number of response categories, then m_j is set to the smaller number of response categories.

medium, and large effect size for Cramér's V to inform our interpretation. Table 7.4 contains Cohen's (1988) values. As can be seen, the V values corresponding to small, medium, and large effect sizes varies as a function of the number of response categories. For instance, with $m_j = 3$ a small or negligible effect would be in the neighborhood of 0.071 (e.g., 0.142 and smaller), a medium or moderate effect would be around 0.212 (e.g., $0.142 < V \leq 0.283$), and a large (strong) effect would be in the vicinity of 0.354 (e.g., $V > 0.283$).

Although most of our values are significant X^2 LDs ($\chi^2_{0.05,-4} = 9.49$), there are three pairs (item pairs 1–2, 2–3, 1–3) that show moderate association and merit closer examination (as discussed above) for the cause(s) of their large values. The remaining significant X^2 LDs reflect small effect sizes (i.e., weak/negligible associations). We would not examine these item pairs further, nor would we examine non-significant X^2 LD(s).

Following our item fit information, we have our item parameter estimates (`coef(. . .)`). As was the case with flexMIRT, we estimate 17 parameters ($m_j * L \delta_{jh}s + \alpha = 2 * 8 + 1 = 17$) that are subsequently rescaled to have $\hat{\alpha} = 1$. For item 1 we have $\hat{\delta}_{11} = -1.299$ and $\hat{\delta}_{12} = -2.060$, for item 2 $\hat{\delta}_{21} = -0.428$ and $\hat{\delta}_{22} = -0.428$, and so on. Our estimated person population variance (`$cov`) is 1.0. Our $\hat{\delta}_{jh}s$ show perfect agreement with those of flexMIRT.

We obtain and display some of our EAP $\hat{\theta}$ s (`fscores(raschpcm, . . .)`) as well as how to save these estimates to an ASCII ("text") file (`write.csv(people_raschpcm, . . .)`) and proceed to our person fit analysis. Because we are working with a Rasch model, our person fit information (`personfit(raschpcm, . . .)`) consists of INFIT, OUTFIT, and their standardized values; although z_h is also produced, we will ignore it. (See Chapter 3 for information on INFIT and OUTFIT.) We use the psych package's `describe` function to obtain descriptive statistics (including skew) on our person fit information. Our maxima for INFIT and OUTFIT are 3.69 and 9.01, respectively. These large maxima indicate that at least one response vector is inconsistent with our model. To get a sense of how many respondents merit closer examination, we obtain the INFIT and OUTFIT distributions by using a dot plot (`ggplot(. . .)`). Figure 7.8 shows that INFIT and OUTFIT are positively skewed, with a few highly dis-

crepant individuals and a few percent with values greater than 2 (INFIT: 2.3%, OUTFIT: 4.1%).

We proceed to examine the response vectors for respondents with INFIT and/or OUTFIT values greater than 2. Our first focus is on individuals whose responses to items located away from their $\hat{\theta}$ s are inconsistent with what is predicted by the model (i.e., large OUTFIT cases). We then examine those individuals whose responses to items

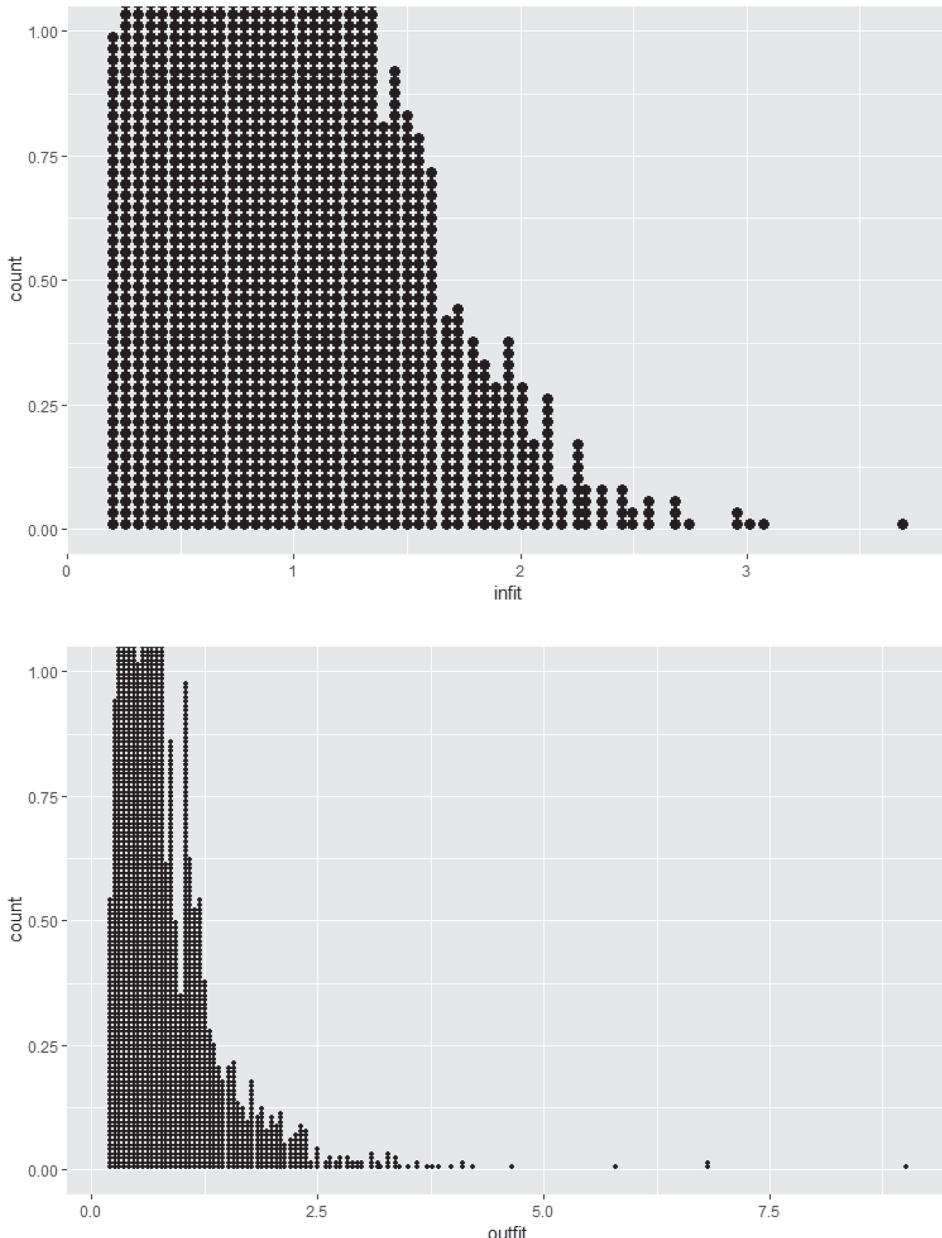


FIGURE 7.8. Dot plot of INFIT (top) and OUTFIT (bottom).

located near their $\hat{\theta}$ are inconsistent with what is predicted by the model (i.e., large INFIT cases). In both phases, we would start with the most extreme valued cases and work toward those cases with less extreme values down to 2.

As an example, let us look at our respondent with an OUTFIT of 9.01. This person has an observed score of 13, $\hat{\theta} = 1.05$ ($s_{\hat{\theta}} = 0.48$), and an INFIT = 2.02. To help our examination of this person's responses, we order the responses by the average item locations.^{12, 13} To this end, we extract our $\hat{\delta}_{jh}$ s (`dEst = coef(raschpcm, . . .)`) and calculate each item's average location (`avgdEst = rowMeans(dEst)`). Because items 6 and 7 are not in order, we sort our responses (`alikedataSort = alikedata[order(avgdEst)]`) to align with ordered item locations. (If our items are originally in order by location, then in our call to `personfit` we could use the argument `stats.only = F` to have the response vectors included in the `personfit` output object.)

To these ordered response data we add the observed score (`alikedataSort$x = rowSums(alikedataSort[,1:8])`), a case identification field (`alikedataSort$id = seq(1,2942,1)`), our fit statistics (e.g., `alikedataSort$outfit = people_raschpcmFit$outfit`), and the $\hat{\theta}$ s with their corresponding standard errors (e.g., `alikedataSort$t_est = people_raschpcm[,1]`). We then sort our data file by OUTFIT to facilitate finding our case. Our case of interest has ordered responses of 02222212. Thus, this individual incorrectly responded to the easiest item, but on more difficult items obtained full credit or partial credit (item 6). Given the preponderance of correct responses, we conjecture that the incorrect response is an anomaly due to, for example, initial inattentiveness. Alternatively, there could be something intrinsic to the item 1 that puts this respondent at a disadvantage; we discuss this issue, differential item functioning (DIF), in Chapter 12. (Because our DIF analysis would have preceded the person fit analysis, this latter conjecture may have been eliminated.) Of course, there are other possible reasons (e.g., copying, distractedness) for the incorrect response that one would also consider and possibly eliminate. Our INFIT is most likely due to receiving full credit on item 7 and partial credit on item 6 (items located around this case's $\hat{\theta} = 1.05$). As such, we do not have reason for concern. For this case, although we recognize that $\hat{\theta}$ may be lower than appropriate, we do not believe that this case degrades our calibration and retain it.

In this fashion, we would proceed with each of the flagged cases looking for, in particular, cases that are inconsistent (i.e., appearing to reflect random guessing) or contradictory (e.g., receiving full/partial credit on items located at the right end of the continuum and no credit on items located at the lower end of the continuum) or cases that show a response set (e.g., with an affective item a pattern such as 01201201). These types of cases do not provide useful information for item parameter estimates and should be removed. Subsequently, the instrument would be recalibrated and the problematic cases' locations re-estimated using the "cleaned" item parameter estimates. Note that when examining polytomous responses that consist of only extreme responses, one should consider the nature of the item. For instance, on a nine-item scale a pattern such as 02020202 may appear at first glance to be of concern. However, this pattern is com-

pletely reasonable if the items are effectively functioning in a dichotomous fashion (e.g., item 1; see Figure 7.4).

The Rating Scale Model

In contrast to proficiency testing, some applications focus on assessing an individual's attitude toward some concept (e.g., immigration), whereas in others one is interested in personality assessment (e.g., social anxiety). These situations tend to use instruments that rely, in part or in whole, on a response format developed by Rensis Likert (1932). This response format typically consists of a series of ordinal categories that range from strongly disagree to strongly agree. This "Likert scale" may contain an even (e.g., 4) or an odd number (e.g., 5 or 7) of response categories. Andrich's (1978b, 1978c; also see Andersen, 1977; Masters, 1982) rating scale (RS) model is appropriate for modeling Likert response scale data as well as performance rating data.

The RS model uses responses from ordered categories and assumes these categories are separated by *thresholds*. Each threshold, τ_h , is on the latent variable's continuum and separates adjacent response categories. For example, assume we are interested in assessing attitudes toward condoms. One of our items is "I prefer to use condoms over other methods of birth control" and uses a four-category Likert response scale. Figure 7.9 shows how the thresholds relate to the response categories. For the moment assume that the item has a location (δ) value of 0. Conceptually, when a person, with location θ , encounters this item the probability of responding in, for example, the "strongly disagree" category or the "disagree" category depends on whether the person is located below or above τ_1 . If $\theta < \tau_1$, then the person responds in the "strongly disagree" category. If the person is located above τ_1 , then following a similar process a response of "disagree" would be determined by whether $\theta < \tau_2$. If $\theta > \tau_2$, then according to this response mechanism a response of "strongly agree" would occur if the person is located above τ_3 ; otherwise the response would be "agree." Thus, the person "passes through" one or more of the thresholds to arrive at their response. The number of thresholds passed is represented by x_j . Therefore, x_j may take on the values from 0 thresholds passed up to and including the mth threshold. When the respondent has passed zero thresholds (i.e., $x_j = 0$), then they remain in the first or lowest category (e.g., strongly disagree). In

I prefer to use condoms over other methods of birth control.

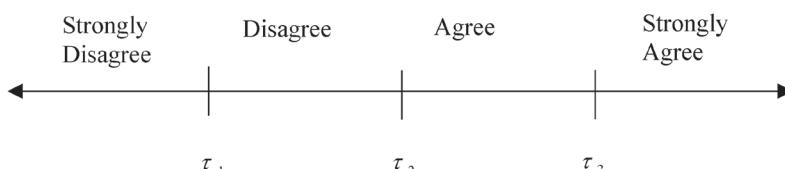


FIGURE 7.9. Representation of item parameter characterization for the RS model (m = 3).

contrast, if the respondent has passed all m thresholds (i.e., $x_j = m$), then they responded in the last or highest category (e.g., strongly agree). Assuming, for instance, that the respondent “agreed” to this example item, then they would have had to pass two thresholds (τ_1 and τ_2) to arrive at the “agree” category and $x_j = 2$.

From Figure 7.9 we see that there is always one fewer thresholds than there are response categories (i.e., there are $m + 1$ response categories). Unlike the PC model in which m is subscripted (i.e., m_j) to reflect that items can differ in terms of the number of operations, with the RS model all items must have the same number of thresholds (i.e., the same number of response categories).¹⁴ These thresholds have the same values for all items. However, this does not mean the thresholds are at the same *locations* on the continuum for all items. Recall that with the Rasch model items on the instrument may have different locations, δ_j s, along the latent variable continuum. The threshold values may be viewed as offsets from an item’s location. As a result, it is the combination of the item’s location and the threshold (offset) value that determines the threshold’s location on the continuum.

As an example, assume we have two items for assessing attitudes toward condom use: “Most of your friends think that condoms are uncomfortable” and “It’s embarrassing to buy condoms,” using the same 4-point Likert scale described above. On a continuum of not favorable to favorable attitudes toward condom use, there is no reason to believe that across a sample of individuals these two items must be located at the same point on the continuum. Figure 7.10 graphically shows the locations of these two items, δ_1 and δ_2 , and how the thresholds for the 4-point Likert item relate to these two item locations. As can be seen, the differences between the τ_h s (i.e., $\tau_1 - \tau_2$, $\tau_1 - \tau_3$, and $\tau_2 - \tau_3$) remain constant across the items. However, the thresholds’ actual locations on the continuum vary as a function of the item’s location (i.e., δ_1 or δ_2). For example, let τ_1, τ_2, τ_3 have the values of $-0.8, -0.2$, and 1.0 , respectively, with items located at $\delta_1 = -1$ and $\delta_2 = 1$. Then the actual location of a threshold h on the continuum for a particular item j would be $\delta_{jh} = \delta_j + \tau_h$. In this example, the first threshold’s location ($h = 1$) on the continuum with respect to item 1 would be $\delta_{11} = \delta_1 + \tau_1 = -1 + (-0.8) = -1.8$, and for the second item the first threshold’s location would be $\delta_{21} = \delta_2 + \tau_1 = 1 + (-0.8) = 0.2$. In a similar fashion, we can determine the locations for the second and third thresholds for these two items.

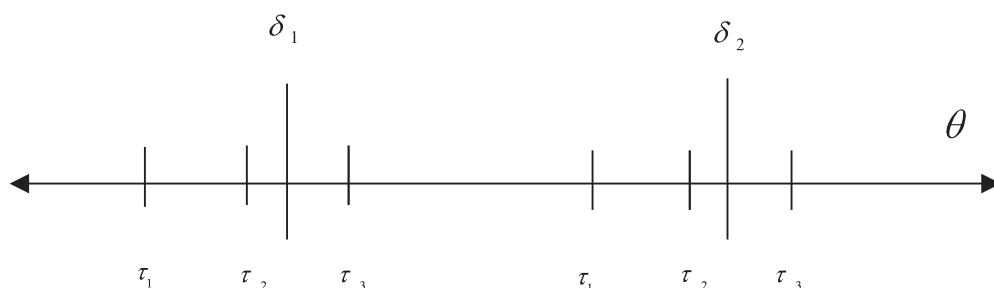


FIGURE 7.10. Graphical representation of item locations and thresholds for two items.

The above ideas and principles may be incorporated into the Rasch model to obtain the RS model

$$p(x_j | \theta, \delta_j, \underline{\tau}) = \frac{\exp\left[\sum_{h=0}^{x_j} (\theta - (\delta_j + \tau_h))\right]}{\sum_{k=0}^m \exp\left[\sum_{h=0}^k (\theta - (\delta_j + \tau_h))\right]} = \frac{\exp\left[-\sum_{h=0}^{x_j} \tau_h + x_j(\theta - \delta_j)\right]}{\sum_{k=0}^m \exp\left[-\sum_{h=0}^k \tau_h + k(\theta - \delta_j)\right]}, \quad (7.2)$$

where $p(x_j | \theta, \delta_j, \underline{\tau})$ is the probability for a person with location θ passing x_j number of thresholds (i.e., responding in a particular category) on an item j located at δ_j with threshold set $\underline{\tau}$ and m is the number of thresholds. As implied above, the number of thresholds and the thresholds themselves are constant across items, so we drop the item subscript j on τ and m (i.e., $\tau_{jh} = \tau_h$, $m_j = m$). The range of x_j is the integer values from 0 to m (i.e., $x_j = \{0, 1, \dots, m\}$). The sum of the thresholds is constrained to be zero, $\sum \tau_h = 0$.¹⁵

Equation 7.2 may be simplified by letting κ_{x_j} represent $-\sum_{h=0}^{x_j} \tau_h$. By substitution of κ_{x_j} into Equation 7.2, one obtains the typical presentation of the RS model (Andrich, 1978b, 1978c)

$$p(x_j | \theta, \delta_j, \underline{\kappa}) = \frac{\exp\left[\kappa_{x_j} + x_j(\theta - \delta_j)\right]}{\sum_{k=0}^m \exp\left[\kappa_k + k(\theta - \delta_j)\right]}, \quad (7.3)$$

where the new term, κ_{x_j} , is referred to as the *category coefficient* and is a function of the τ_j s. By definition, $\kappa_{x_j} = 0$ when x_j is zero; otherwise $\kappa_{x_j} = -\sum_{h=1}^{x_j} \tau_h$, with x_j taking on a value from 1 up to the m th threshold.

In the context of the RS model, the person's location (θ) may be interpreted as an individual's attitude, and the item's location (δ_j) may be interpreted as the item's affective value (Andrich, 1978c) or the difficulty of endorsing the item. Given the relationship between Equations 7.2 and 7.3, it may be evident that the RS model is an extension of the Rasch model. Therefore, the model's underlying assumptions are a unidimensional latent space and items' similar capacity to discriminate among respondents; there is also an assumption that there is equal discrimination at the thresholds. When the RS model is applied to dichotomous data, the RS model reduces to the Rasch model (i.e., there is one threshold, $m = 1$, $\tau_1 = \delta_j$). For convenience p_{xj} is used in lieu of $p(x_j | \theta, \delta_j, \underline{\kappa})$ in the following.

As is the case with the PC model, each response category has an ORF. Moreover, the ORFs for any item always consist of at least one monotonically nondecreasing ORF and one monotonically nonincreasing ORF. There is typically a unimodal ORF for each additional response category. Figures 7.11 and 7.12 contain example ORFs for two items. These two items use a four-category Likert response scale (strongly disagree, disagree, agree, strongly agree) and come from an instrument designed to measure attitude toward condom use. Such an instrument might be used, for example, as part of an HIV awareness program. Item 1 asks the respondents whether condoms offer good protection and item 2 asks if they are embarrassed to purchase condoms. Assume that item 1 is located

at ($\delta_1 = -0.98$) and that item 2 has a δ_3 of 0.70. For these two items, as well as all the other items on the instrument, assume the thresholds have the values of $\tau_1 = -0.30$, $\tau_2 = -0.02$, and $\tau_3 = 0.32$. Using these values with the model in Equation 7.3 produced the ORFs in Figures 7.11 and 7.12. As can be seen, each item has a monotonically nondecreasing ORF and a monotonically nonincreasing ORF representing the strongly agree and strongly disagree response categories, respectively. The agree and disagree response categories are reflected in ORFs that are unimodal and symmetric.¹⁶

The ORFs in Figure 7.11 may be interpreted as indicating that individuals in the narrow band between -1.28 and -1.0 are likely to disagree with this item. Similarly, respondents in the range -1.0 to -0.67 are likely to agree with the item. Outside this range of approximately six-tenths of a logit (i.e., from -0.67 to -1.28), respondents will, by and large, have a higher probability of strongly disagreeing with this item if their attitudes are below -1.28 or strongly agreeing with this item if their attitudes are above -0.67 .

Comparing Figure 7.11 with Figure 7.12 shows the same pattern, albeit at a different point along the continuum. If the right side of the continuum reflects people who

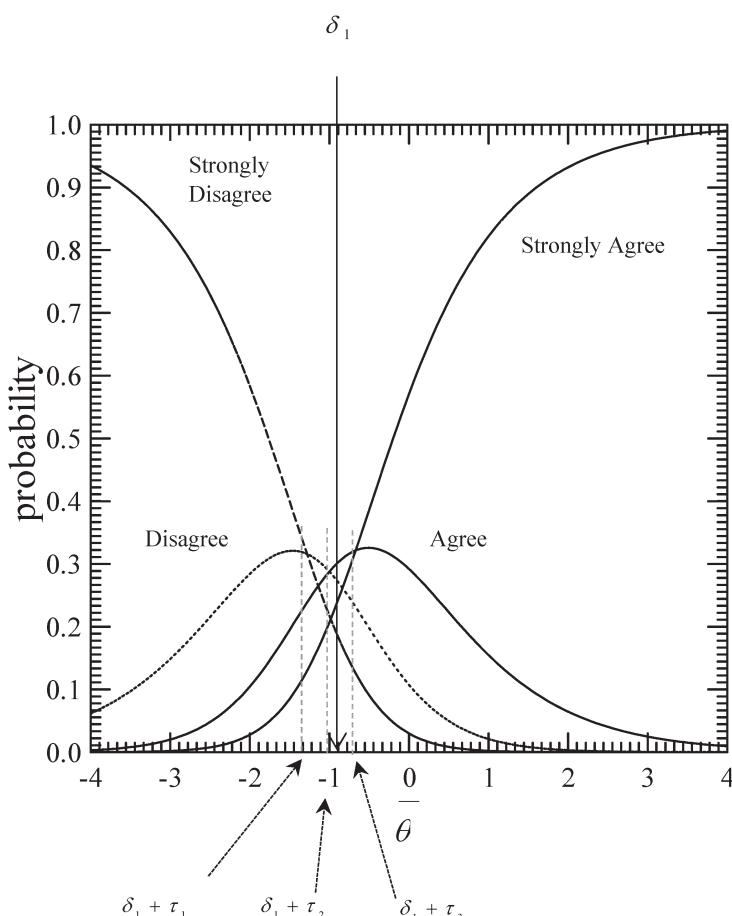


FIGURE 7.11. RS model ORFs for a four-category Likert item with $\delta_1 = -0.98$.

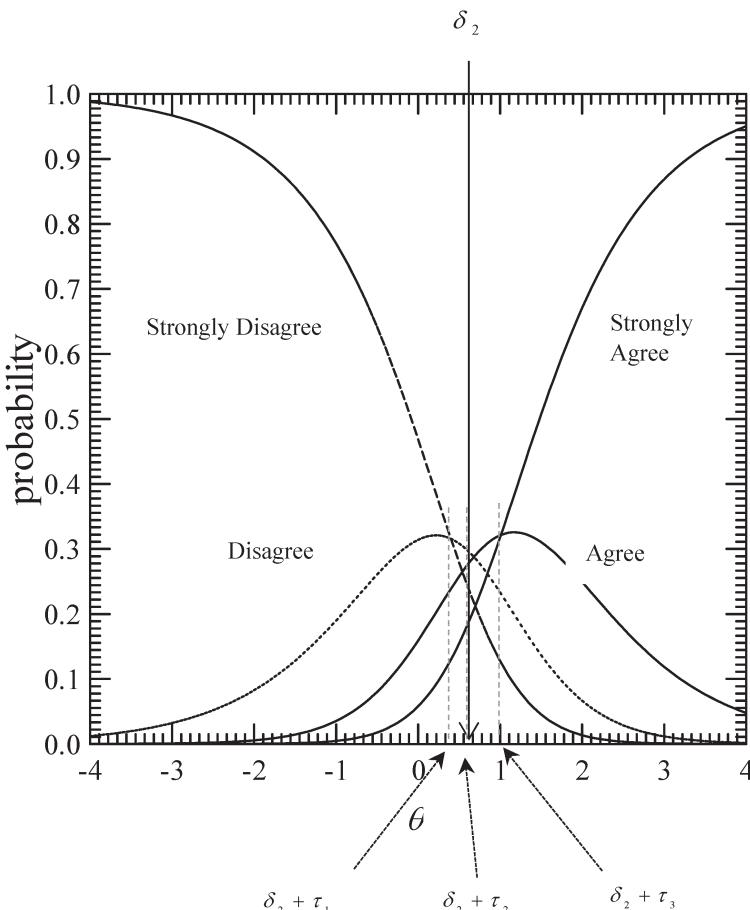


FIGURE 7.12. RS model ORFs for a four-category Likert item with $= 0.70$.

have negative attitudes toward condom use, then item 2 (“It’s embarrassing to buy condoms”) is an item that individuals who have relatively negative attitudes toward condom use (e.g., located above 1.0) will tend to strongly agree with or to strongly endorse. In contrast, a comparatively large portion of the continuum reflects individuals who are comfortable in purchasing condoms, even though some of them have an average/neutral or somewhat positive attitudes toward condom use.

Similarly, given item 1’s location, respondents who are somewhat below average, as well as those who are positive in their attitude toward condom use, will tend to strongly agree that condoms offer good protection. Overall, and given the threshold values, the nature of this item set is that there is a tendency to either strongly disagree or strongly agree with an item. That is, this item set has a strong tendency to polarize individuals, and the four response categories behave almost, but not quite, as two categories.

One might ask, “Where on the continuum does the probability of strongly disagreeing equal the probability of disagreeing with item 1?” (This point is represented

by the leftmost vertical dash line in Figure 7.11). In other words, what is the location of τ_1 , given item 1's location? To determine the threshold's location on the continuum, one simply adds the threshold's value to the item's location. Therefore, we have $\delta_1 + \tau_1 = -0.98 + (-0.30) = -1.28$ as the point of intersection of the strongly disagree ORF and the disagree ORF. With respect to this item, the second and third threshold locations would be located at -1.0 and -0.67 , respectively. In a similar fashion, we may determine the thresholds' locations on the continuum for the third item (Figure 7.12).

Figure 7.11 shows that the relative locations of the thresholds (i.e., with respect to item 1's location) correspond to the transitions between the ORFs of adjacent categories. This is identical to the PC model's transition locations interpretation. In terms of the PC model notation, the actual location of threshold 1 for item 1 would be $\delta_{11} = \delta_1 + \tau_1$ or in general, $\delta_{jh} = \delta_j + \tau_h$. Masters (1982) has shown that his PC model subsumes the RS model; also see Masters and Wright (1984) as well as Wright and Masters (1982). Moreover, under certain constraints, the RS model is a special case of Bock's nominal response model; the nominal response model is discussed in Chapter 9. As is the case with the PC model, the sum of the probabilities across response categories for a fixed θ equals 1.0.¹⁷

Conceptual Parameter Estimation of the RS Model

The formulas for the RS model estimation may be found in Andrich (1978c). As is the case with the PC and Rasch models, the principle of maximizing a likelihood function is used. Two commonly used approaches to obtain the parameter estimates are JMLE (UCON) (see Chapter 3) and MMLE (see Chapter 4). For estimation purposes, a person's observed score (i.e., the simple unweighted sum across items of the number of thresholds passed) is a sufficient statistic for estimating the person's location on the continuum. As is the case with the other members of the Rasch family, individuals who obtain the same observed score have the same location on the continuum. Furthermore, the item score (i.e., the unweighted sum of responses across people) is a sufficient statistic for estimating item j 's location, δ_j . As a result, items that have the same item score have the same location. With respect to the category coefficient, the total number of responses with respect to all respondents and all items that are associated with category x is a sufficient statistic for estimating the corresponding κ_x (Andrich, 1978c). One implication of a common set of thresholds across items is that the thresholds need only be estimated once for an item set.

Example: Application of the RS Model to an Attitudes Toward Condoms Scale, JMLE, BIGSTEPS

The example's data come from the Voluntary HIV Counseling and Testing Efficacy Study performed by the Center for AIDS Prevention Studies (2003). This study was concerned with the effectiveness of HIV counseling and testing for the prevention of new HIV

infections. As part of this study, respondents were surveyed about their attitudes toward condoms. Six items from the survey are used in this example. The six items are statements people had made about condoms. Respondents were asked how much they agreed with each of the statements on a 4-point response scale (1 = “strongly disagree,” 2 = “disagree more than I agree,” 3 = “agree more than I disagree,” 4 = “strongly agree”).¹⁸ Given the (“negative”) wording of the statements, a respondent who strongly agreed with a statement was indicating a less favorable attitude toward condom use. We assume that we have evidence supporting the tenability of a unidimensional latent space. This evidence might be obtained, for example, through the factor analysis of a polychoric item correlation matrix; SAS’s plcorr keyword with proc freq can be used to obtain these coefficients. As mentioned above, the validity of the measures provides additional information concerning this assumption.

As we did in Chapter 3, we initially use BIGSTEPS and then the R package mixRasch to also calibrate our data; both programs use JMLE. (Other programs that could be used are WINSTEPS [JMLE], ConQuest [JMLE, MMLE], flexMIRT, IRTPRO, PARSCALE, and mirt.) Because this example builds on the program’s features that are introduced in Chapter 3, we now focus on new aspects of the output. (The author’s website shows the analysis of these data using PARSCALE and Endnotes 19 and 20 present flexMIRT analyses.)

Table 7.5 contains the command file for specifying the BIGSTEPS calibration of the Attitude Toward Condoms scale. The layout of this file is similar to that shown in Chapter 3; however, with this example, we introduce user-specified item labeling. Specifically, the items’ text is paraphrased to be used as the items’ labels. The CODES line specifies the observed response values of 1 through 4, and MODELS=R specifies our model.

Table 7.6 shows that PROX and UCON converged; PROX iterated 4 times and UCON iterated 19 times. By looking at the PROX iteration history, we see that 219 respondents were dropped from the analysis. Inspection of the data shows that 119 of these individuals responded “strongly disagree” to all items (i.e., all 1s) and the remaining 100 responded “strongly agree” to each item (i.e., all 4s).

Because our calibration involves a polytomous model, the CONVERGENCE TABLE contains information not seen in Chapter 3 (cf. Table 3.3). For instance, the Max Logit Change columns now contain information not only about persons (MEASURES), but also about items (STEPS). This additional item information is due to the model’s transition locations. In a well-behaved situation, the values in both MEASURES and STEPS should decrease with increasing iterations. In this example, both of these show this decrease. The CATEGORY RESIDUAL and STEP CHANGE columns indicate the difference between the observed and expected count for any category and the maximum logit change, respectively. The CATEGORY RESIDUAL and STEP CHANGE are expected, as they do here, to decrease across iterations. In addition to information indicating the person (CASE) and item (ITEM) that are farthest from meeting the convergence criterion, we have information about the ordinal position of the response category, CAT, that is farthest from meeting the criterion. As stated in Chapter 3, if the standardized residuals have a mean close to zero with a standard deviation (S.D.) of approximately 1.0, then this indicates the data are following Rasch model assumptions. If the M and SD

TABLE 7.5. BIGSTEPS Command File for RS Model Calibration of the Attitude Towards Condoms Scale

```

;RS Calibration of Attitude Towards Condoms data
&INST
TITLE='Attitude Towards Condoms'
NI=6
ITEM1=5
CATEGS=4
XWIDE=1
CODES=1234
NCOLS=10
MISSING=8
MODELS=R
CURVES=111
CATREF=0
STBIAS=Y
TABLES=1110011001001000100000
NORMAL=Y
NAME1=1
PERSON=CASE           ← person labela
ITEM=item              ← general item label
PFILE=HIV00COM.PF
IFILE=HIV00COM.IF
&END
EMBARRASS BUY CONDOM      ← item label for item 1
CONDOM NOT GOOD FEEL      ← item label for item 2
EMBARRASS TO PUT ON CONDOM   :
CONDOMS BREAK/SLIP OFF      :
PARTNER WANTS CONDOM CHEAT    :
FRIENDS CONDOM UNCOMFORTABLE   ← item label for item 6
END NAMES

```

^aThe text following the ← is provided to help the reader understand the corresponding input.

are substantially different from 0 and 1, respectively, then there is a departure from the assumption that randomness is normally distributed and this affects the fit statistics.

Table 7.7 shows BIGSTEPS's Variable Map table. Recall that the Variable Map (Chapter 3) shows how the distributions of respondents and items relate to one another. Because persons and items are located on the same continuum, the term MEASURE refers to either $\hat{\theta}$ (left side) or $\hat{\delta}$ (right side). The numerical demarcations are logits. The left side of the leftmost panel shows the distribution of respondents (symbolized by either "#" or a "."); at the bottom of the table, the legend indicates that each "#" represents 30 individuals, and each "." indicates from 1 to 29 persons). Respondents who were likely to "strongly agree" to the items are located toward the top of this distribution, whereas those who were likely to "strongly disagree" are located toward the bottom of the table's left panel. Given the negative wording of the items, the top represents a less favorable attitude toward condom use and the bottom reflects a positive attitude toward condom use. As can be seen, the majority of individuals tend to fall between 0.0 and the positive attitude toward condom use end of the continuum.

The next three panels show how the items' (transition) locations relate to the person

TABLE 7.6. Abridged Program Control Parameters (Table 0.1) and Iteration History Table (Table 0.2)

TABLE 0.1 Attitude Towards Condoms

```

TITLE= Attitude Towards Condoms
CONTROL FILE: condom.con
OUTPUT FILE: condom.lis

CONTROL VARIABLES:
Input Data Format      PAIRED = N          Item Delete/Anchor
  DATA =                 REALSE = N          IDFILE =
  NAME1 = 1              STBIAS = Y          IDELQU = N
NAMLEN = 4              -----               IAFILE =
ITEM1 = 5               Misfit Selection   IANCHQ = N
ITLEN = 30              FITI = 2.000       -----
NI = 6                  FITP = 2.000       Person Delete/Anchor
XWIDE = 1               OUTFIT = Y          PDFILE =
INUMB = N               LOCAL = N          PDELQU = N
:
PERSON = CASE           -----
ASCII = Y               Convergence Control RFILE =
-----                   MPROX = 10        SFILE =
User Scaling            MUCON = 0          XFILE =
  UMEAN = .000          LCONV = .010       -----
USCALE = 1.000          RCONV = .500       Data Reformat
UDECIM = 2              TARGET = N         FORMAT =
UANCH = Y               -----
Scale Structure          GROUPS =          GRPFRM = N
ADJUSTMENT              MODELS = R         KEYFRM = 0
EXTRSC = .500           STKEEP = N         MODFRM = N
HIADJ = .250             -----
LOWADJ = .250           -----
3473 CASE    Records Input
-----
```

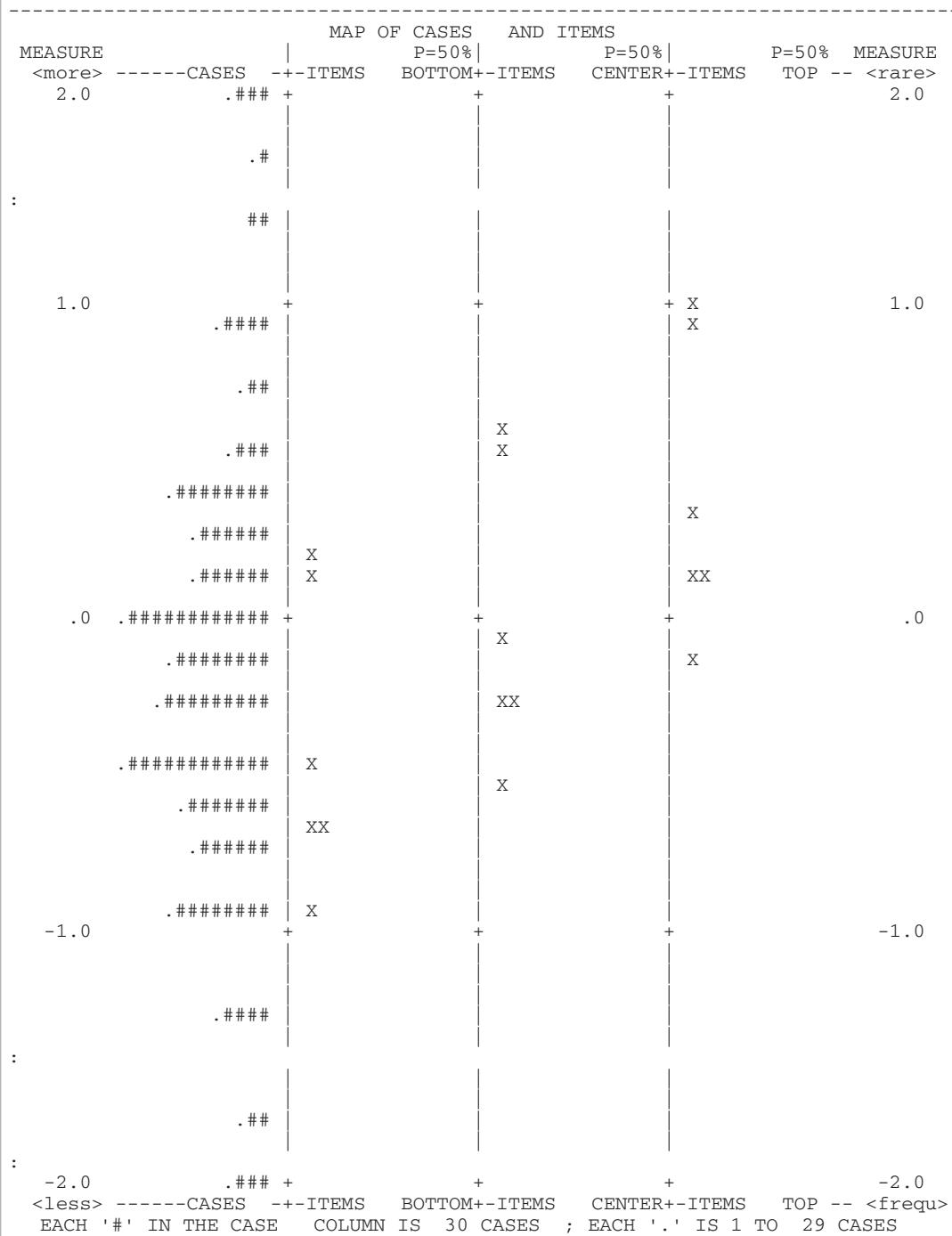
CONVERGENCE TABLE

PROX ITERATION	ACTIVE CASES	COUNT ITEMS	CATS	EXTREME CASES	5 RANGE ITEMS	MAX LOGIT MEASURES	CHANGE STEPS
1	3473	6	5	4.23	.73	3.1355	1.1028
2	3373	6	4	5.86	1.16	-2.0426	-.5026
3	3254	6	4	6.15	1.28	-.1464	-.1740
4	3254	6	4	6.26	1.30	-.0519	-.0291
UCON ITERATION	MAX SCORE RESIDUAL*	MAX LOGIT CHANGE	LEAST CASE	CONVERGED ITEM	CATEGORY CAT	STEP RESIDUAL	CHANGE
1	543.93	.9585	3	3*	1	-445.05	.1709
2	367.56	.2833	562	3*	4	366.01	.0854
3	257.23	-.1632	37	3*	4	388.41	.1430
4	217.48	-.1456	18	3*	4	216.10	-.0637
5	157.73	-.0920	18	3*	4	149.73	-.0423
6	108.77	-.0597	18	3*	4	99.81	-.0274
7	72.83	-.0386	18	3*	4	64.54	-.0177
8	47.98	-.0249	18	3*	4	40.99	.0115
9	31.30	-.0160	18	3*	4	25.59	.0079
10	20.17	-.0101	18	3*	4	16.19	.0059
11	12.91	-.0063	18	3*	4	10.30	.0038
12	8.20	.0040	695	3*	4	6.87	.0020
13	5.33	-.0026	18	3*	4	4.66	-.0013
14	3.49	-.0018	18	3*	4	3.13	-.0009
15	2.23	-.0012	18	3*	4	1.86	-.0005
16	1.43	-.0007	18	3*	4	1.08	.0003
17	.92	-.0005	18	3*	4	.85	.0002
18	.59	-.0003	18	3*	4	.53	.0002
19	.39	-.0002	18	3*	4	.31	.0001

Standardized Residuals N(0,1) Mean: -.02 S.D.: 1.01

TABLE 7.7. RS (BIGSTEPS) Item-Person Map for the Attitude Towards Condoms Scale

TABLE 1.1 Attitude Towards Condoms FIFTEEN SUBPARTS
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS



distribution. Each item's location is symbolized by an X. The first item panel (labeled ITEMS BOTTOM) "locates" the items by using the inflection point of the IRF corresponding to the bottommost category (i.e., the "strongly disagree" category), the second item panel (labeled ITEMS CENTER) shows the location of the items at the center of the rating scale (i.e., where responding in the top and bottom categories is equally probable [Linacre & Wright, 2001]), and the third (rightmost) item panel (labeled ITEMS TOP) "locates" the items by using the inflection point of the IRF of the topmost category (i.e., "strongly agree"); the location of the inflection point (i.e., $p = 0.5$) does not necessarily correspond to δ_{jh} . As a result, with the RS model the topmost X in each panel reflects the same item. Similarly, the second to the top X in each panel represents another item and so on for the remaining four items on the scale. For instance, the topmost X in each panel represents item 3, the second to the topmost X in each panel represents item 2, and the bottommost X in each panel reflects item 6; identifying which X represents which item comes from the output discussed below. Figure 7.13 shows how these X locations for the ITEMS BOTTOM and ITEMS TOP relate to item 6's ORFs. The ITEMS BOTTOM and ITEMS TOP locations are sometimes called category boundaries (Masters, 1982).

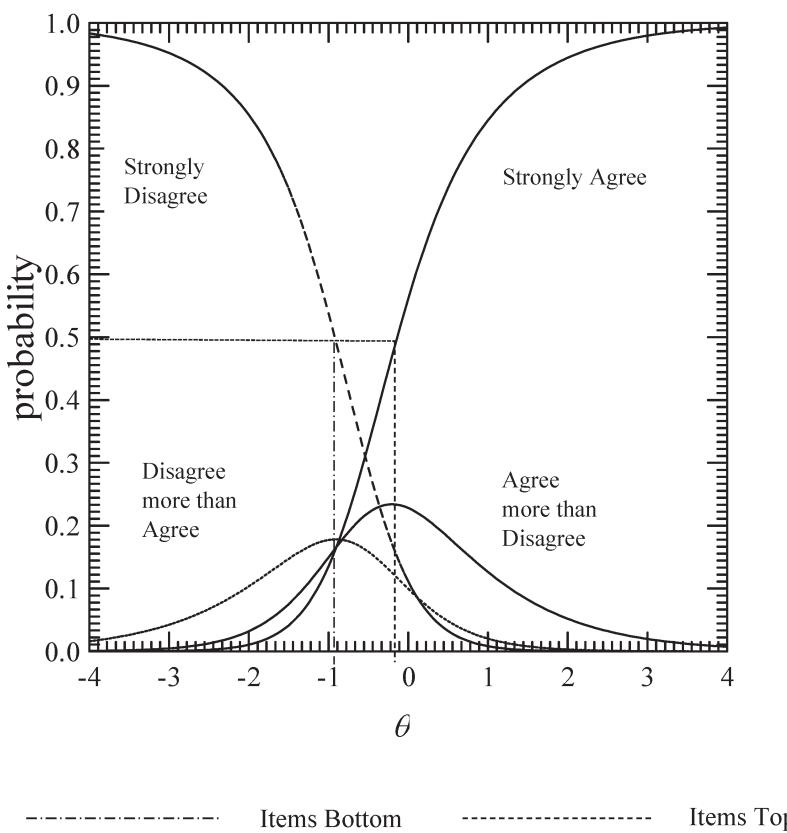


FIGURE 7.13. RS model ORFs for item 6 with ITEMS BOTTOM and ITEMS TOP locations identified ($\hat{\delta}_6 = -0.54$ with $\hat{\tau}_1 = 0.65$, $\hat{\tau}_2 = -0.29$, and $\hat{\tau}_3 = -0.36$).

By comparing the item locations to the person distribution, we have an idea of how well the items are functioning in measuring the people. For example, we see from the lowest X in the rightmost panel that roughly half the respondents (i.e., people located slightly below 0.0 and above) have a probability of at least 0.5 of responding in the “strongly agree” category for item 6. Conversely, there are more than 270 persons with locations of about -1 or less who have a probability of at least 0.5 of responding in the item’s “strongly disagree” category (i.e., the lowest location X in the first item panel). Table 7.7 shows that, given the distribution of respondents, the items (collectively) cover most of the attitudinal range where people are located. Moreover, it may be anticipated that this range of attitude will facilitate the estimation of the item δ_{jh} s. In short, this table not only allows one to see how well the respondents’ distribution matches the range of the instrument, but also provides an idea of how well the items are distributed. Using this information, one may anticipate where on the θ continuum one may experience greater difficulty in estimating person as well as item locations.

Table 7.8 (BIGSTEPS’ TABLE 3.1) shows descriptive statistics for the calibration as well as information about how well the test separates the respondents’ measures. Recall from our discussion in Chapter 3 that the top half of this table presents summary information on the respondents (i.e., SUMMARY OF . . . CASES), whereas the bottom half presents summary information on the items (i.e., SUMMARY OF . . . ITEMS).

The descriptive statistics on the observed score (labeled RAW SCORE) shows the mean observed score for the 3254 NON-EXTREME respondents is 14.1, with a range from 7 to 23. (The observed score, X , is the sum of the item responses. With six four-category items the range of X is from 6 to 24 (i.e., $6*1 = 6$ and $6*4 = 24$, where 1 = strongly disagree, 4 = strongly agree.) We know from the iteration history (see Table 7.6) that only 3254 of the 3473 respondents are used in the calibration. We now see that of the 3473 respondents 100 obtained the MAXIMUM EXTREME SCORE of 24 and 119 persons obtained the MINIMUM EXTREME SCORE of 6; these 219 respondents are considered EXTREME CASES. The columns labeled MEASURE and MODEL ERROR refer to $\hat{\theta}$ and its standard error, respectively. Therefore, across the 3254 persons, the mean person location estimate is -0.16 with a standard deviation of 0.64, and the smallest and the largest $\hat{\theta}$ s are -1.71 and 1.78, respectively. In general, individuals tend to be slightly favorably predisposed toward condom use.

We would like to see the INFIT MNSQ and OUTFIT MNSQ values be close to 1 (or 0 for the INFIT ZSTD and OUTFIT ZSTD).²¹ At the model level, we have model fit with respect to persons. Recall that the REAL RMSE is the root mean squared error calculated on the basis that misfit in the data is due to departures in the data from model specifications, whereas MODEL RMSE is the same statistic but is calculated on the basis that the data fit the model. In the person half of TABLE 3.1, the RMSEs are calculated across people. Small RMSEs indicate good model-data fit situations with respect to people.

The person SEPARATION index is the ratio of person variability to the RMSE (i.e., SEPARATION = ADJ.SD/RMSE); as such, the index is in standard error units. In addition, the person RELIABILITY index is equivalent to Cronbach’s alpha (Linacre & Wright, 2001), with values close to or equal to 1 considered to be good. This example’s

TABLE 7.8. RS (BIGSTEPS) Item Location Estimates for the Attitude Towards Condoms Scale

TABLE 10.1 Attitude Towards Condoms
 INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

ITEMS STATISTICS: MISFIT ORDER												
ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTBIS		ITEMS	
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.			
5	8444	3254	-.25	.02	1.02	4.9	1.12	4.5	A .23	PARTNER WANTS CONDOM CHEAT		
4	8585	3254	-.29	.02	1.02	.9	1.11	3.9	B .17	CONDOMS BREAK/SLIP OFF		
6	9465	3254	-.54	.02	.95	-2.4	1.05	1.7	C .21	FRIENDS CONDOM UNCOMFORTABLE		
1	7773	3254	-.06	.02	.97	-1.3	.99	-.4	c .27	CONDOM NOT GOOD FEEL		
3	5573	3254	.62	.02	.98	-.9	.91	-2.2	b .34	EMBARRASS TO PUT ON CONDOM		
2	5897	3254	.51	.02	.97	-1.2	.89	-2.8	a .34	EMBARRASS BUY CONDOM		
MEAN	7623.	3254.	.00	.02	1.00	.0	1.01	.8				
S.D.	1426.	0.	.42	.00	.05	2.4	.09	2.8				

TABLE 10.2 Attitude Towards Condoms
 INPUT: 3473 CASES. 6 ITEMS ANALYZED: 3254 CASES. 6 ITEMS. 4 CATS

ITEMS FIT GRAPH: MISFIT ORDER													
ENTRY NUMBER	MEASURE		INFIT MEAN-SQUARE					OUTFIT MEAN-SQUARE					ITEMS
	-	+	0	0.7	1	1.3	2	0	0.7	1	1.3	2	
5	*		:	*	:		A	:	*	:			PARTNER WANTS CONDOM CHEAT
4	*		:	*	:		B	:	*	:			CONDOMS BREAK/SLIP OFF
6	*		:	*	:		C	:	*	:			FRIENDS CONDOM UNCOMFORTABLE
1	*		:	*	:		C	:	*	.			CONDOM NOT GOOD FEEL
3		*	:	*	:		b	:	*	.			EMBARRASS TO PUT ON CONDOM
2		*	:	*	:		a	:	*	.			EMBARRASS BUY CONDOM

TABLE 10.3 Attitude Towards Condoms
 INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

TABLE 10.4 Attitude Towards Condoms
 INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

MOST MISFITTING RESPONSE STRINGS									
ITEM	OUTMNSQ	CASE							
		33333333333333332222222222111	33221	22211					
		3044443322221097665543110920960041494110446553	42						
		85430076988620802636164849072224178924957646970224							
		05515233886758140573163768306787671441040776979582							
		high							
5	PARTNER WANTS	1.12 A	.11..11..1..1..1111..11..111..1..1..						
4	CONDOMS BREAK	1.11 B1..11..1..1..1..1..1..						
6	FRIENDS CONDO	1.05 C	22..1..1..1..						
3	EMBARRASS TO	.91 b)						
2	EMBARRASS BUY	.89 a						
									low
		33333333333333332222222222111963322194222116553242							
		304444332222109766554311092022041424110446970524							
		85430076988620802636164849072784178941957646979	82						
		05515233886758140573163768306	76714	04077					

(continued)

TABLE 7.8. (continued)

TABLE 10.5 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

MOST UNEXPECTED RESPONSES		CASE
ITEM	MEASURE	
		333333333333332222222222111 33221 22211
		3044443322221097665543110920960041494110446553 42
		85430076988620802636164849072224178924957646970224
		05515233886758140573163768306787671441040776979582
		high-----
6 FRIENDS CONDO	-.54 C	22..1..1..1.....1.1111.....
4 CONDOMS BREAK	-.29 B1..11....1..1..1.1.....
5 PARTNER WANTS	-.25 A	.11.11..1..1..1111.11.111..1..
2 EMBARRASS BUY	.51 a4444.....4..4.33
3 EMBARRASS TO	.62 b44444.44..4..
		-----low
		3333333333333322222222211963322194222116553242
		3044443322221097665543110920220041424110446970524
		85430076988620802636164849072784178941957646979 82
		05515233886758140573163768306 76714 04077

TABLE 13.1 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

ITEMS STATISTICS: MEASURE ORDER									
ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	OUTFIT MNSQ	PTBIS ZSTD	ITEMS	PTBIS CORR.
3	5573	3254	.62	.02	.98	-.9	.91	-2.2	.34
2	5897	3254	.51	.02	.97	-1.2	.89	-2.8	.34
1	7773	3254	-.06	.02	.97	-1.3	.99	-.4	.27
5	8444	3254	-.25	.02	1.10	4.9	1.12	4.5	.23
4	8585	3254	-.29	.02	1.02	.9	1.11	3.9	.17
6	9465	3254	-.54	.02	.95	-2.4	1.05	1.7	.21
MEAN	7623.	3254.	.00	.02	1.00	.0	1.01	.8	
S.D.	1426.	0.	.42	.00	.05	2.4	.09	2.8	

person RELIABILITY of 0.44 indicates a low level of consistency in the ordering of person location estimates across different attitudes toward condom scales. More information on the SEPARATION and RELIABILITY indices is found in Chapter 3 and Appendix G, "The Separation and Reliability Indices."

The RMSE, ADJ. SD, SEPARATION, and RELIABILITY statistics are repeated using all 3473 respondents in the subsection entitled WITH . . . EXTREME CASES. These results are ignored because they are affected by the number of extreme respondents and the value used in their estimation (i.e., the EXTRSC estimation adjustment criterion).

The bottom half of TABLE 3.1 (SUMMARY OF . . . ITEMS) contains the same statistical indices as used for describing the respondents but is now focused on the items. The columns labeled MEASURE and MODEL ERROR refer to the $\hat{\delta}$ s and their standard errors. As would be expected from a program that uses item centering, the mean $\hat{\delta}$ is 0 with a standard deviation of 0.42. The minimum $\hat{\delta}$ is -0.54 and the maximum $\hat{\delta}$ is 0.62. Although not indicated in this table, in Table 7.10 (BIGSTEPS' TABLE 10.1) we see that the minimum $\hat{\delta}$ is for item 6, whereas the maximum is item 3's $\hat{\delta}$. The results indicate that, overall, there is model-data fit from an item perspective (INFIT MNSQ and OUTFIT MNSQ); that there is little error in the estimation (REAL RMSE and MODEL RMSE); and that we have a good item RELIABILITY value indicating the instrument is creating a well-defined variable.

Table 7.9 contains item threshold-level calibration information and summary information about the response categories. The CATEGORY LABEL column contains the CODES labels, where 1 = “strongly disagree,” 2 = “disagree more than I agree,” 3 = “agree more than I disagree,” and 4 = “strongly agree.” For each category, the number of persons responding in each category across items is presented in the OBSERVED COUNT column (e.g., for CATEGORY LABEL 1, the OBSERVED COUNT is the sum of the frequency of 1 across all 6 items). The AVERAGE MEASURE and its EXPected value are defined in the table’s legend, where B_n is the person location estimate, $\hat{\theta}$, and D_i is the item location estimates, $\hat{\delta}_s$. This AVERAGE MEASURE column shows that, on average, as one progresses from “strongly disagree” to “strongly agree,” there is an increase in the respondents’ (less positive) attitudes toward condom use.

The COHERENCE EXP% and COHERENCE OBS% are ratios focused on the degree of agreement between what is observed for a response category and what would be expected for that category. In this situation, 100% represents the best-case scenario, and values less than 50% are considered to be “inferential insecure” (Linacre & Wright, 2001). These ratios differ from one another only in their respective denominators, with the numerators reflecting the number of times an observation and its expectation fell within the same response category. For COHERENCE EXP% a response category’s count is divided by all *expectations* in the response category, whereas for COHERENCE OBS% a response category’s count is divided by all *observations* in the response category. Linacre and Wright (2001) state that COHERENCE EXP% assesses the extent to which measures corresponding to a response category predict ratings in it. In contrast, COHERENCE OBS% is concerned with the extent to which ratings in a response category predict measures corresponding to it. It can be seen that the “strongly disagree” and “strongly agree” response categories do a good job of predicting the ratings within them (COHERENCE EXP%), whereas the “disagree more than I agree” and “agree more than I disagree” ratings do comparatively better in predicting the measures corresponding to a response category (COHERENCE OBS%) than vice versa. The INFIT MNSQ and OUTFIT MNSQ values for the items are relatively close to or equal to their expected value of 1, indicating consistency between the data and the RS model. That is, a common set of thresholds (τ_h s) for the entire item set appears to be appropriate.

The STEP CALIBRATION column contains the estimated transition location for the category of interest *relative* to the transition location from the category below the one of interest. For example, assume the category of interest is “disagree more than I agree.” Therefore, the transition from “strongly disagree” to “disagree more than I agree” occurs at the relative position of 0.65; the lowest category, 1 (“strongly disagree”), has no *prior* transition location and is shown as NONE. The transition from “disagree more than I agree” to “agree more than I disagree” occurs at the relative position of -0.29, and the last transition from “agree more than I disagree” to “strongly agree” occurs at the relative position of -0.36. Therefore, the τ estimates are $\tau_1 = 0.65$, $\tau_2 = -0.29$, and $\tau_3 = -0.36$, with a $s_e(\hat{\tau}_h)$ of 0.02 for each. To determine the actual, not relative, location on the continuum of the transition from, say, “strongly disagree” to “disagree more than I agree” for a particular item, would require also knowing the item’s location; this is demonstrated below. As we would expect, the sum of the $\hat{\tau}$ s is 0.0. As we see, the differences between successive $\hat{\tau}$ s are not equal (i.e., $|\hat{\tau}_1 - \hat{\tau}_2| \neq |\hat{\tau}_2 - \hat{\tau}_3|$).

TABLE 7.9. RS (BIGSTEPS) Calibration Summary Results for the Attitude Towards Condoms Scale

TABLE 3.2 Attitude Towards Condoms

INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

SUMMARY OF MEASURED STEPS

CATEGORY LABEL	OBSERVED COUNT	AVERAGE EXP.		COHERENCE		INFIT OUTFIT		STEP CALIBRATN
		MEASURE		EXP%	OBS%	MNSQ	MNSQ	
1	8004	-.67	-.65	85%	39%	.99	1.07	NONE
2	2608	-.24	-.29	18%	52%	1.00	.99	.65
3	3131	.13	.07	23%	48%	.86	.83	-.29*
4	5781	.41	.44	71%	22%	1.04	1.03	-.36*

AVERAGE MEASURE is mean of (Bn-Di), EXP. is expected value.

AVERAGE MEASURE IS mean of (B1-B7), EXP. IS expected value.
 $\text{EXP\%} = (\text{expected \& observed}) / (\text{all expected})$ [MEASURE->RATING?]

EXP% = (expected & observed) / (all expected)
OBS% = (expected & observed) / (all observed)

CATEGORY LABEL	STEP CALIBRATN	STEP S.E.	SCORE-TO-MEASURE			THURSTONE THRESHOLD
			AT	CAT.	----ZONE----	
1	NONE		(-1.31)	-INF	.85	
2	.65	.02	-.37	-.85	-.02	.38
3	-.29	.02	.33	-.02	.84	-.05
4	-.36	.02	(1.34)	.84	+INF	.40

CATEGORY PROBABILITIES: MODES - Step measures at intersections

The figure is a scatter plot with the following characteristics:

- X-axis:** Labeled "CASE" at the bottom center, ranging from -2 to 2.
- Y-axis:** Labeled "[MINUS]" on the left side, ranging from .0 to 1.0.
- Data Points:** Represented by '+' symbols.
- MEASURE Values:** The plot shows data for various MEASURE values, with distinct patterns for each:
 - MEASURE = 0.0:** Vertical column of '+' symbols at CASE = 0.
 - MEASURE = 0.2:** Horizontal row of '+' symbols at [MINUS] = 0.2.
 - MEASURE = 0.4:** Horizontal row of '+' symbols at [MINUS] = 0.4.
 - MEASURE = 0.5:** Horizontal row of '+' symbols at [MINUS] = 0.5.
 - MEASURE = 0.6:** Horizontal row of '+' symbols at [MINUS] = 0.6.
 - MEASURE = 0.8:** Horizontal row of '+' symbols at [MINUS] = 0.8.
 - MEASURE = 1.0:** Vertical column of '+' symbols at [MINUS] = 1.0.
- Other Patterns:** There are also diagonal and curved patterns of '+' symbols, particularly around CASE = 0 and [MINUS] = 0.2 to 0.8.

The τ s are expected to increase across the response categories and when there is a reversal in their order; as with categories 3 and 4, they are flagged with an “*.” A reversal indicates that the corresponding category is not as likely to be chosen as the other categories. In this case, given the number of reversals relative to the number of response categories, these items are behaving primarily in a dichotomous, not polytomous, fashion. Given these $\hat{\tau}$ s, it is not surprising that the CATEGORY PROBABILITIES plot shows that persons responding to these items are primarily using just the “strongly disagree” and “strongly agree” response categories. For completeness, the THURSTONE THRESHOLD indicates the location of the median probability. That is, at these locations, the probability of observing the categories below equals the probability of observing the categories equal to or above. This is the point on the variable at which the category interval begins (Linacre & Wright, 2001).

The $\hat{\delta}$ s are presented in BIGSTEPS' TABLE 10.1:ITEMS STATISTICS:MISFIT ORDER (see Table 7.10); BIGSTEPS' TABLE 13.1 presents the same information as TABLE 10.1 but in descending order of $\hat{\delta}$. Because the output is presented in misfit (MNSQ) order, we need to examine the ENTRY NUMBR column to determine the item number for a given $\hat{\delta}_j$ (labeled MEASURE). As can be seen, item 1 is located at $\hat{\delta}_1 = -0.06$ with an $s_e(\hat{\delta}_h) = 0.02$, item 2 is located at $(\hat{\delta}_2 =) 0.51$ with an $s_e(\hat{\delta}_2) = 0.02$, and so on. As we would expect from an item-centering approach, the mean $\hat{\delta}$ is (approximately) 0.0.

By combining the $\hat{\tau}$ s from BIGSTEPS' TABLE 3.2 with the $\hat{\delta}$ s, we can calculate the $\hat{\delta}_{jh}$ s as well as produce the items' ORFs. For example, for item 1 the transition from “strongly disagree” to “disagree more than I agree” is

$$\hat{\delta}_{11} = \hat{\delta}_1 + \hat{\tau}_1 = -0.06 + 0.65 = 0.59,$$

the transition from “disagree more than I agree” to “agree more than I disagree” occurs at

$$\hat{\delta}_{12} = \hat{\delta}_1 + \hat{\tau}_2 = -0.06 + (-0.29) = -0.35,$$

and the transition from “agree more than I disagree” to “strongly disagree” has a location for item 1 at

$$\hat{\delta}_{13} = \hat{\delta}_1 + \hat{\tau}_3 = -0.06 + (-0.36) = -0.42.$$

Figure 7.14 contains the ORFs for item 1 based on these intersection points. As can be seen, the interpretation of the CATEGORY PROBABILITIES from TABLE 3.2 applies to the ORFs for item 1; this is also the interpretation for all the items.

The items' INFIT MNSQ and OUTFIT MNSQ values are provided in this table, although the corresponding graphical presentation (TABLE 10.2) is easier to interpret. The interpretation of these indices is discussed in Chapter 3. From the information presented in TABLE 10.1 and TABLE 10.2, we conclude the items are behaving in a fashion consistent with the model.

With polytomous data, it is useful to produce TABLE 10.3 because it provides a

TABLE 7.10. RS (BIGSTEPS) Item Location Estimates for the Attitude Towards Condoms Scale

TABLE 10.1 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

ITEMS STATISTICS: MISFIT ORDER										
ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTBIS CORR.	ITEMS
5	8444	3254	-.25	.02	1.10	4.9	1.12	4.5	A .23	PARTNER WANTS CONDOM CHEAT
4	8585	3254	-.29	.02	1.02	9.1	1.11	3.9	B .17	CONDOMS BREAK/SLIP OFF
6	9465	3254	-.54	.02	.95	-2.4	1.05	1.7	C .21	FRIENDS CONDOM UNCOMFORTABLE
1	7773	3254	-.06	.02	.97	-1.3	.99	-.4	c .27	CONDOM NOT GOOD FEEL
3	5573	3254	.62	.02	.98	-.9	.91	-2.2	b .34	EMBARRASS TO PUT ON CONDOM
2	5897	3254	.51	.02	.97	-1.2	.89	-2.8	a .34	EMBARRASS BUY CONDOM
MEAN	7623.	3254.	.00	.02	1.00	.0	1.01	.8		
S.D.	1426.	0.	.42	.00	.05	2.4	.09	2.8		

TABLE 10.2 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

ITEMS FIT GRAPH: MISFIT ORDER										
ENTRY NUMBR	MEASURE	INFIT MEAN-SQUARE	OUTFIT MEAN-SQUARE	ITEMS						
5	*	: * : 0	A : * : 0	PARTNER WANTS CONDOM CHEAT						
4	*	: * : 0	B : * : 0	CONDOMS BREAK/SLIP OFF						
6	*	: * : 0	C : * : 0	FRIENDS CONDOM UNCOMFORTABLE						
1	*	: * : 0	c : * : 0	CONDOM NOT GOOD FEEL						
3	*	: * : 0	b : * : 0	EMBARRASS TO PUT ON CONDOM						
2	*	: * : 0	a : * : 0	EMBARRASS BUY CONDOM						

TABLE 10.3 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

ITEMS OPTION/DISTRACTOR FREQUENCIES: MISFIT ORDER										
NUM	NONMISS	MISSING	R% SCR	1 % SCR	2 % SCR	3 % SCR				
		4	% SCR							
5A	3473	0	0 **	1243	35	1 380	10	2 440	12	3
		1410	40	4						
4B	3473	0	0 **	956	27	1 596	17	2 728	20	3
		1193	34	4						
6C	3473	0	0 **	744	21	1 451	12	2 774	22	3
		1504	43	4						
1D	3473	0	0 **	1338	38	1 492	14	2 602	17	3
		1041	29	4						
3E	3473	0	0 **	2280	65	1 344	9	2 272	7	3
		577	16	4						
2F	3473	0	0 **	2157	62	1 345	9	2 315	9	3
		656	18	4						

TABLE 10.4 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

MOST MISFITTING RESPONSE STRINGS		CASE
ITEM	OUTMNSQ	
		3333333333333332222222222111 33221 22211
		3044443222221097665543110920960041494110446553 42
		85430076988620802636164849072224178924957646970224
		05515233886758140573163768306787671441040776979582
		high-
5 PARTNER WANTS	1.12 A	..11.11..1.1..1111.11.111.1..1.....
4 CONDOMS BREAK	1.11 B1..11....1..1..1.1.....
6 FRIENDS CONDO	1.05 C	22..1..1..1.....1.1111.....
3 EMBARRASS TO	.91 b44444.44.4..
2 EMBARRASS BUY	.89 a4444....4..4.33
		-----low
		3333333333333332222222222111963322194222116553242
		304444322222109766554311092022041424110446970524
		85430076988620802636164849072784178941957646979 82
		05515233886758140573163768306 76714 04077

(continued)

TABLE 7.10. (continued)

TABLE 10.5 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

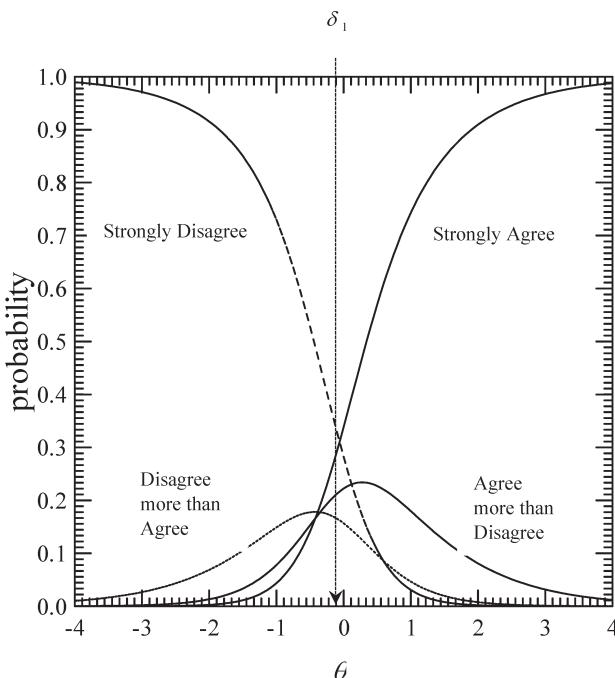
MOST UNEXPECTED RESPONSES

ITEM	MEASURE	CASE
6 FRIENDS CONDO	.54 C	32211 22211
4 CONDOMS BREAK	-.29 B	3044443322221097665543110920960041494110446553 42
5 PARTNER WANTS	-.25 A	85430076988620802636164849072224178924957646970224
2 EMBARRASS BUY	.51 a	05515233886758140573163768306787671441040776979582
3 EMBARRASS TO	.62 b	high----- 3333333333333222222222111 33221 22211 3044443322221097665543110920960041494110446553 42 85430076988620802636164849072224178924957646970224 05515233886758140573163768306787671441040776979582
		-----low 3333333333333222222222111963322194222116553242 3044443322221097665543110920220041424110446970524 85430076988620802636164849072784178941957646979 82 05515233886758140573163768306 76714 04077

TABLE 13.1 Attitude Towards Condoms
INPUT: 3473 CASES, 6 ITEMS ANALYZED: 3254 CASES, 6 ITEMS, 4 CATS

ITEMS STATISTICS: MEASURE ORDER

ENTRY	RAW	COUNT	MEASURE	ERROR	INFIT	OUTFIT	PTBIS	ITEMS
NUMBER	SCORE		MNSQ	ZSTD	MNSQ	ZSTD	CORR.	
3	5573	3254	.62	.02	.98	-.9	.91	-2.2
2	5897	3254	.51	.02	.97	-1.2	.89	-2.8
1	7773	3254	-.06	.02	.97	-1.3	.99	-.4
5	8444	3254	-.25	.02	1.10	4.9	1.12	4.5
4	8585	3254	-.29	.02	1.02	.9	1.11	3.9
6	9465	3254	-.54	.02	.95	-2.4	1.05	1.7
MEAN	7623.	3254.	.00	.02	1.00	.0	1.01	.8
S.D.	1426.	0.	.42	.00	.05	2.4	.09	2.8

**FIGURE 7.14.** RS model ORFs for item 1 ($\hat{\delta}_1 = -0.06$ with $\hat{\tau}_1 = 0.65$, $\hat{\tau}_2 = -0.29$, and $\hat{\tau}_3 = -0.36$).

breakdown of the respondents' use of each response category for each item. This information is useful for diagnosing estimation problems and/or for identifying items that should be rewritten. As is the case with TABLES 10.1–10.4, the items are listed in order of their misfit. This table's layout has the item number first (NUM) followed by the number of respondents to the item (NONMISS). The next three columns contain the number of omissions (MISSING), the percentage of omission (R%), and the code (SCR) used to represent missing. The remaining columns present the category frequency, the percentage of respondents in the category, and the code representing the category response for each of the categories. In this example, the information "wraps around," and the fourth response category is listed in the second panel (i.e., MISSING R% SCR) that contains the missing data information. This is why below the MISSING R% SCR label we find the fourth response category's 4% SCR label.

Interpreting BIGSTEPS' TABLE 10.3, we have that the first item listed is item 5, there are 3473 respondents to this item with 0 MISSING, and the percentage missing (R% =) is 0%; the “**” indicates that no code is used to represent missing. Of the 3473 respondents, 1243 (% = 35%) responded in the first category (i.e., SCR = 1), 380 (10%) responded in category 2, 440 (12%) responded in category 3, and, wrapping around, 1410 or 40% responded in the fourth category (“strongly agree”). Therefore, each item presented actually occupies two lines in the table. The next item presented is item 4, and its information is listed on the third and fourth lines of the table's body. This pattern is continued for the remaining items.

As we can see, each response category attracted at least 272 respondents. As such, the thresholds should be well estimated. In general, the first and fourth response categories had the largest number of respondents regardless of the item. The only exception to this occurred with item 6 and the “agree more than I disagree” response (category 3). From these frequencies we see that each response category is being used. However, from Figure 7.14 we know the items are primarily functioning in a dichotomous fashion.

If we have fit problems with one or more items, then BIGSTEPS' TABLE 10.4 and TABLE 10.5 can be useful for diagnostic purposes (TABLE 10.4 differs from TABLE 10.5 in that TABLE 10.4 contains the items in terms of their OUTFIT values, whereas TABLE 10.5 contains the items in terms of their locations). As an example, in TABLE 10.4 the first case listed has the id of #3380 (one reads across the rows within a column of the CASE panel to determine the id code). We know from the person estimates that this respondent is estimated to be located at $\hat{\theta}_{3380} = 1.26$ (OUTFIT = 3.66); to save space, the estimated person locations output is not shown. This person's response of 2 is the most misfitting response to item 6. Item 6 had an OUTFIT of 1.05 value and is estimated to be located at -0.54 . Therefore, given the item and person estimated locations, it is expected that this person would have provided a response of 3 or 4. That is, it is expected that this person, who is located toward the “less favorably predisposed toward condom use” end of the continuum, would have agreed or strongly agreed with the item that condoms are uncomfortable. Instead, this person indicated they disagreed more than agreed with the item. The transition location estimate for “strongly agree” on this item is $\hat{\delta}_{63} = \hat{\delta}_6 + \hat{\tau}_3 = -0.54 - 0.36 = -0.90$, and this is more than 2.1 logits

($-90 - 1.26 = -2.16$) below the person's estimated location. (The next individual listed, #3055, has the same estimated location as #3380.)

The last person in the table, #242, is estimated to be located at $\hat{\theta}_{242} = -1.24$ ($\text{OUTFIT} = 3.38$). This person is located toward the "more favorably predisposed toward condom use" end of the continuum. However, they provided a response of 3 ("agree more than disagree") to item 2 ($\hat{\delta}_2 = 0.51$), "It's embarrassing to buy condoms." That is, given that this respondent is located more than two logits below the transition location estimate of $(\hat{\delta})_{21}$ of 1.16, we expect they would *not* select the "agree more than disagree" response. Rather, this person is expected to "strongly disagree" or "disagree" more than "agree" with this item (i.e., select category 1 or 2).

As discussed above, our model–data fit analysis should include examinations of the invariance and conditional independence assumptions. The invariance procedure presented in Chapter 3 would be applied to the $\hat{\delta}_{jh}s$. In terms of assessing the conditional independence assumption, one could use Q_3 . With the RS model, the residual for an item is the difference between an individual's observed response and the individual's expected (category) response on the item. In our example, the observed responses are 1 through 4, and the expected response is given by the weighted sum of the response category probabilities according to the RS model, that is

$$\mathcal{E}(x_j | \hat{\theta}_i) = \sum_{k=1}^{m+1} kp(x_j | \hat{\theta}_i).$$

If a 0-based counting system is used (e.g., Dodd, 1990), then

$$\mathcal{E}(x_j | \hat{\theta}_i) = \sum_{k=0}^m kp(x_j | \hat{\theta}_i).$$

The Pearson correlation coefficient would be applied to the residuals for items j and z ; $d_{ij} = x_{ij} - \varepsilon(x_j | \hat{\theta}_i)$ and $d_{iz} = x_{iz} - \varepsilon(x_z | \hat{\theta}_i)$.

Once the researcher achieves satisfactory model–data fit with a scale, the IRT advantages over traditional approaches, such as the capacity to estimate a person's attitude free of a scale's characteristics, are available to the researcher as part of their study. In fact, subsequent administrations of the scale do not necessarily require that the new response data be recalibrated to obtain person attitudinal estimates. Instead, one may take advantage of working with the Rasch family of models to create a concordance table using the satisfactory calibration results (i.e., all individuals with the same observed score receive the same $\hat{\theta}$). This table would contain the possible observed scores and their corresponding $\hat{\theta}s$. In this way, subsequent administrations would only require determining the respondent's observed score and then using the table to determine the corresponding $\hat{\theta}$.

Example: Application of the PC Model to an Attitudes Toward Condoms Scale, JMLE, mixRasch

In Chapter 3, we introduced the R package `mixRasch` (Willse, 2011, 2014). As above, we assume the relevant libraries (`mixRasch`, `Hmisc`) are loaded into our R workspace;

Table 7.11 shows our session. (Endnote 22 presents the `mirt` analysis.) Our data file is an ASCII file that uses a comma to separate values (csv) and has been annotated to have format information on its first line. To input these data, we use `read.table` function with the `skip` argument to skip the first line (`skip = 1`) and the `sep` argument to specify the symbol (i.e., `sep = ","`) used to delimit our values. Our data reside in the data frame (object) `condomsdata`. We verify that our data were correctly read by using the `head` and `tail` functions; our sample size is 3473. Moreover, we verify that all responses codes for our six items fall between 1 and 4 (inclusive) by using the `describe(condomsdata[, seq(2,7)])` function from the `Hmisc` package. By using the `seq` function to specify columns 2–7 of `condomsdata`, we omit obtaining our `id` variable (i.e., column 1). As can be seen, none of our response categories have very small or zero frequencies.

`mixRasch` (and `mirt`) assumes the data frame passed to it contains only the response data to be calibrated. As such, we remove our `id` variable from `condomsdata` (i.e., `within(condomsdata, rm(id))`) prior to calling `mixRasch`. Our call to `mixRasch` specifies the Rating scale model with three thresholds `mixRasch(. . . , steps = 3, model = "RSM", . . .)` as our calibration model. Moreover, to obtain pseudo $\hat{\theta}$ s for zero variance response vectors, we use the `treat.extreme` argument (`treat.extreme = 0.5`). Thus, 0.5 will be subtracted from the perfect score and 0.5 added to a zero score. (These zero variance response vectors are not used for estimating the item parameters.) By default, `mixRasch` will terminate after a maximum of 50 iterations. Because our calibration required only 15 iterations, we know we have a converged solution; our call to `getEstDetails(. . .)` confirms this (`$convergeFlag` is `TRUE`). From `raschRSM$info.fit`, we see that 3254 (`$N.persons`) of the 3473 were used in the calibration, with the difference comprising respondents with zero variance response vectors ($X = 0$ or $X = 24$).

Our information criteria (`$AIC`, `$BIC`, `$CAIC`) and log likelihood (`$loglik`) value (after being multiplied by -2) could be used for model comparisons. At the item level we graphically examine our `INFIT` and `OUTFIT` values by using the `itemFitPlot(. . .)` function (Figure 7.15). The shaded portions indicate `INFIT` or `OUTFIT` values outside of the 0.7–1.3 range. Because these `INFIT` and `OUTFIT` values correspond to those of `BIGSTEPS`, the interpretation given above applies here. Using the common cutoff points of 0.7 and 1.3, there does not appear to be any apparent serious fit problems because the fit statistics are within acceptable bounds. (Please recall our comments about common cutoffs in our discussion of Table 3.4.)

Displaying the contents of our output object `raschRSM` shows our item location estimates (e.g., $\hat{\delta}_1 = -0.062$, $\hat{\delta}_2 = 0.505$, . . . , $\hat{\delta}_6 = -0.535$), and their corresponding standard errors (e.g., $s_e(\hat{\delta}_1) = 0.017$, $s_e(\hat{\delta}_2) = 0.019$, . . . , $s_e(\hat{\delta}_6) = 0.017$), `INFIT` (e.g., item 1: 0.974), `OUTFIT` (e.g., item 1: 0.988) and the standardized values (e.g., item 1: `in.Z = -1.320` and `out.Z = -0.413`). Our estimated thresholds are given in the `step1`, `step2`, and `step3` columns (i.e., $\hat{\tau}_1 = 0.653$, $\hat{\tau}_2 = -0.290$, and $\hat{\tau}_3 = -0.363$, with a $s_e(\hat{\tau}_1)$ of 0.017, and $s_e(\hat{\tau}_2) = s_e(\hat{\tau}_3) = 0.018$). These values agree perfectly with those of `BIGSTEPS`. As shown above, we can determine the transition point from one category to the next category for each item by $\delta_{jh} = \delta_j + \tau_h$. Examination of the traditional item discrimination

TABLE 7.11. mixRasch Session for the Rasch RS Model Calibration of the Attitudes Towards Condoms Scale

```

> # load mixRasch and Hmisc
> # data file is comma separated values, skip file format info on first line, & add
> #   variable names; could have used read.csv(..) instead of read.table(..)
> condomsdata = read.table(file.choose(), sep=",", skip=1, col.names=c("id",paste0("I",1:6)))

> head(condomsdata,5)
  id I1 I2 I3 I4 I5 I6
1 1 1 1 1 1 1
2 2 1 1 1 1 1
3 3 1 1 1 4 1
4 4 1 1 1 1 1
5 5 1 1 1 1 1

> tail(condomsdata,5)
  id I1 I2 I3 I4 I5 I6
3469 3469 4 4 4 4 4
3470 3470 4 4 4 4 4
3471 3471 4 4 4 4 4
3472 3472 4 4 4 4 4
3473 3473 4 4 4 4 4

> frequency distribution and stats on just items (not case id)
> Hmisc::describe(condomsdata[,seq(2,7)])

```

Variables 3473 Observations						
I1	n	missing	distinct	Info	Mean	
3473	0	4	0.908	2.388	1.392	
Value	1	2	3	4		
Frequency	1338	492	602	1041		
Proportion	0.385	0.142	0.173	0.300		
I2	n	missing	distinct	Info	Mean	Gmd
3473	0	4	0.752	1.847	1.18	
Value	1	2	3	4		
Frequency	2157	345	315	656		
Proportion	0.621	0.099	0.091	0.189		
I3	n	missing	distinct	Info	Mean	Gmd
3473	0	4	0.711	1.754	1.098	
Value	1	2	3	4		
Frequency	2280	344	272	577		
Proportion	0.656	0.099	0.078	0.166		
I4	n	missing	distinct	Info	Mean	Gmd
3473	0	4	0.924	2.621	1.345	
Value	1	2	3	4		
Frequency	956	596	728	1193		
Proportion	0.275	0.172	0.210	0.344		
I5	n	missing	distinct	Info	Mean	Gmd
3473	0	4	0.884	2.581	1.44	

(continued)

TABLE 7.11. (*continued*)

Value	1	2	3	4		
Frequency	1243	380	440	1410		
Proportion	0.358	0.109	0.127	0.406		
<hr/>						
I6	n	missing	distinct	Info	Mean	Gmd
	3473	0	4	0.896	2.875	1.279
<hr/>						
Value	1	2	3	4		
Frequency	744	451	774	1504		
Proportion	0.214	0.130	0.223	0.433		
<hr/>						
> condomsdata=within(condomsdata,rm(id))	# remove id from data frame					
> raschRSM=mixRasch(condomsdata,steps=3,model="RSM",info.fit=T,treat.extreme = 0.5)						
Iteration: 1, Largest Parameter Change: 1.535976						
Iteration: 2, Largest Parameter Change: 0.3575107						
Iteration: 3, Largest Parameter Change: 0.09410605						
Iteration: 4, Largest Parameter Change: 0.04252516						
Iteration: 5, Largest Parameter Change: 0.02473229						
Iteration: 6, Largest Parameter Change: 0.01508276						
Iteration: 7, Largest Parameter Change: 0.01093306						
Iteration: 8, Largest Parameter Change: 0.007941255						
Iteration: 9, Largest Parameter Change: 0.005754092						
Iteration: 10, Largest Parameter Change: 0.004161427						
Iteration: 11, Largest Parameter Change: 0.003005409						
Iteration: 12, Largest Parameter Change: 0.002168382						
Iteration: 13, Largest Parameter Change: 0.0015634						
Iteration: 14, Largest Parameter Change: 0.001126681						
Iteration: 15, Largest Parameter Change: 0.0008116992						
> getEstDetails(raschRSM)						
\$model						
[1] "RSM"						
\$nC						
[1] 1						
\$iter						
[1] 15						
\$maxChange						
[1] 0.0008116992						
\$convergeFlag						
[1] TRUE						
\$runTime						
Time difference of 4.86423 secs						
> raschRSM\$info.fit						
\$AIC						
[1] 41408.9						
\$BIC						
[1] 41555						
\$CAIC						
[1] 41579						
\$loglik						
[1] -20680.45						
\$N parms						
[1] 24						

(continued)

TABLE 7.11. (*continued*)

```

$N.persons
[1] 3254

> itemFitPlot(raschRSM, fitStat="infit", colTheme="greys")
> itemFitPlot(raschRSM, fitStat="outfit", colTheme="greys") # produces Figure 7.15

> raschRSM
    difficulty      se infit   in.Z outfit   out.Z
  I1     -0.062 0.017 0.974 -1.320  0.988 -0.413
  I2      0.505 0.019 0.971 -1.167  0.894 -2.880
  I3      0.624 0.020 0.976 -0.879  0.910 -2.224
  I4     -0.285 0.017 1.016  0.845  1.106  3.724
  I5     -0.246 0.017 1.096  4.795  1.123  4.328
  I6     -0.535 0.017 0.949 -2.447  1.052  1.674

    step1    se1 step2    se2 step3    se3
I1 0.653 0.017 -0.29 0.018 -0.363 0.018
I2 0.653 0.017 -0.29 0.018 -0.363 0.018
I3 0.653 0.017 -0.29 0.018 -0.363 0.018
I4 0.653 0.017 -0.29 0.018 -0.363 0.018
I5 0.653 0.017 -0.29 0.018 -0.363 0.018
I6 0.653 0.017 -0.29 0.018 -0.363 0.018

Time difference of 4.86423 secs

Number of iterations: 15

> raschRSM$item.par$itemDescriptives
    itemMean      pBis      bis
  I1 2.387561 0.5830477 0.7127619
  I2 1.847394 0.6222150 0.7813155
  I3 1.754103 0.6200452 0.7791664
  I4 2.621365 0.5292464 0.6329342
  I5 2.580766 0.5634588 0.7133480
  I6 2.874748 0.5493796 0.6701187

> head(raschRSM$person.par,6)
    theta SE.theta r infit   in.Z outfit   out.Z
  1 -2.218417 1.2419195 NA      NA      NA      NA
  2 -2.218417 1.2419195 NA      NA      NA      NA
  3 -0.962917 0.4801878 3 1.674440 5.6166335 1.326626 2.1735520
  4 -2.218417 1.2419195 NA      NA      NA      NA
  5 -2.218417 1.2419195 NA      NA      NA      NA
  6 -1.238027 0.5794214 2 1.221548 0.7141603 1.240682 0.6548134

> tail(raschRSM$person.par,4)
    theta SE.theta r infit in.Z outfit out.Z
  3470 2.352038 1.312461 NA      NA      NA      NA
  3471 2.352038 1.312461 NA      NA      NA      NA
  3472 2.352038 1.312461 NA      NA      NA      NA
  3473 2.352038 1.312461 NA      NA      NA      NA

> # produces Figure 7.16
> hist(raschRSM$person.par$infit,main="Person Infit Distr",xlab="infit")
> hist(raschRSM$person.par$outfit,main="Person Outfit Distr",xlab="outfit")

> # save person MLE to external file
> write.csv(raschRSM$person.par, file = "peopleRaschRSM.csv")

```

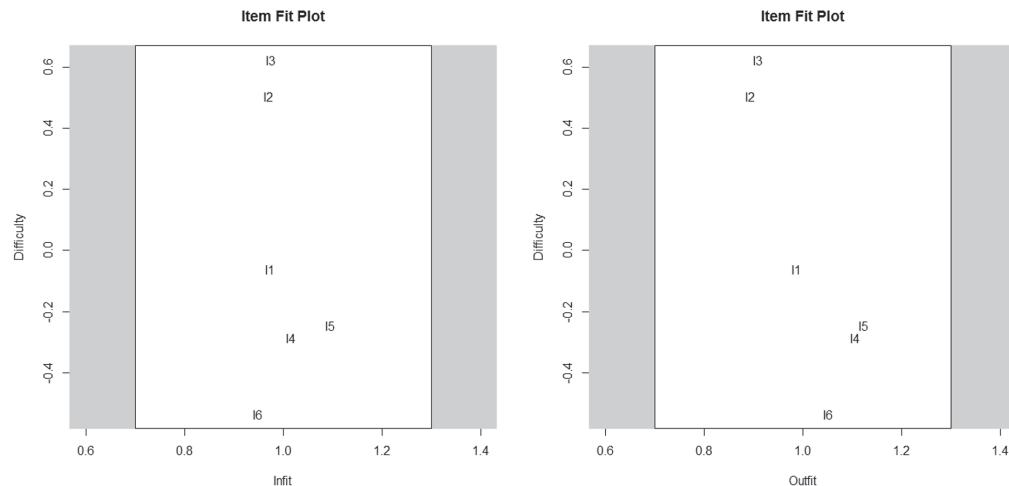


FIGURE 7.15. Item INFIT (left) and OUTFIT (right) plots.

indices point biserial (`pBis`) and biserial (`bis`) shows little variability across items. As shown in Chapter 3, we can obtain a variable map using the `personItemPlot` function.

Because `mixRasch` implements JMLE, there is no additional step to obtain our $\hat{\theta}$ s. These estimates are found in our output objects' `person.par` variable. The first six and last four respondents are shown. The first two cases ($\hat{\theta} = -2.218$) are pseudo $\hat{\theta}$ s for $X = 0$. As such, the fit statistics are calculated and are shown as NA (not available). Our third and sixth cases have had their MLE location estimates -0.963 and -1.238 , respectively. Similarly, our last four cases are all pseudo $\hat{\theta}$ s for $X = 24$. As above, we would proceed to examine our person fit. To facilitate this examination, we obtain histograms (`hist(. . .)`) of our person INFIT and OUTFIT values (Figure 7.16). There are clearly some respondents with large OUTFIT (and to a lesser extent INFIT) values whose response vectors are not consistent with the model. We would proceed to analyze these response vectors as we did above with our PC calibration using `mirt`.

How Large a Calibration Sample?

In contrast to the dichotomous models, the polytomous PC and RS models contain (potentially) more item parameters to estimate. Given these additional item parameters, one might expect the sample size recommendations for dichotomous models to not generalize to polytomous models.

Walker-Bartnick (1990) studied the stability of PC model item parameter estimates using MSTEPS. (MSTEPS [Wright, Congdon, & Shultz, 1988] is a precursor to BIGSTEPS and uses JMLE.) She found a ratio of at least 2 persons per item parameter (i.e., minimum 2:1) to produce stable item and person parameter estimates, regardless

of the number of response categories. Choi, Cook, and Dodd (1997) investigated the parameter recovery for the PC model for instruments of varying lengths and two levels of the number of response categories for an item. Their calibration software, MULTILOG (Thissen, Chen, & Bock, 2003), uses MMLE. They suggested the sample size ratio could be a more complete guideline if it took into account the number of transition locations per item. That is, although they obtained accurate estimation with a sample size of 250, they recommended that, for a given number of total parameters, the sample size to number of item parameters ratio should be larger if there are a large number of response categories than if there are a small number of categories.

With respect to the RS model, French and Dodd (1999) studied how well the item and person parameters were estimated under various sample sizes, θ distributions, and δ distributions. Although they kept the instrument length fixed at 30 items, they varied the number of response categories ($m = 4$ and $m = 5$). They used PARSCALE (MMLE) for parameter estimation. Consistent with the findings of others (e.g., Reise & Yu, 1990), they found that sample size does not appear to impact the recovery of person locations (i.e., when MMLE is used for estimation). Moreover, with a minimum sample size ratio of approximately 1.8:1 PARSCALE was able to recover the item parameters very well.

Given the foregoing, then, for example, a 30-item instrument with a four-category response scale and using the 2:1 ratio and JMLE, we would expect stable results with 180 people (30 items * 3 transition location parameters * 2 = 180). Assuming the respondents distribute themselves uniformly across the response categories, then a 2:1 ratio would lead to the expectation that each category would have approximately 45 respondents (180 people/4 categories = 45). However, a uniform distribution of individuals across response categories is not likely to occur in practice. Therefore, with smaller samples, it is more likely that certain response categories would not have any, or would have a relatively small number of, respondents, and, as such, their parameter estimates would

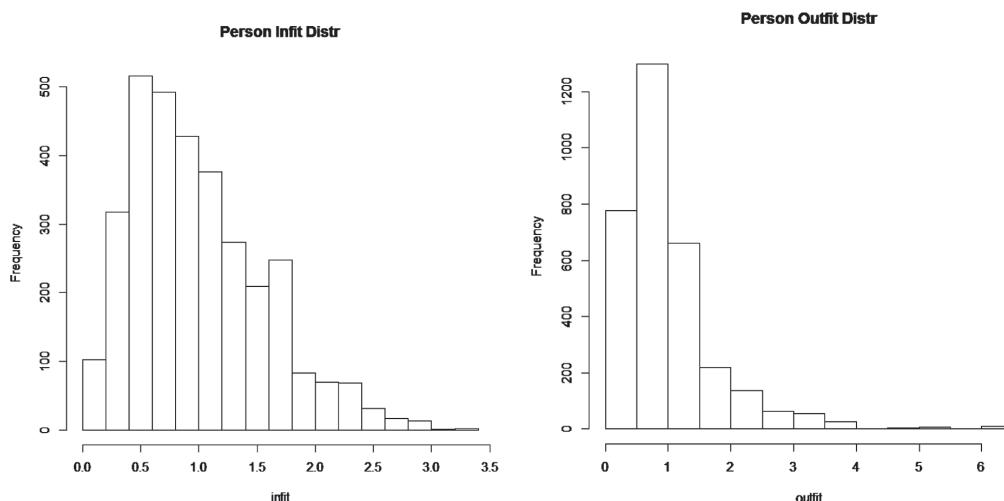


FIGURE 7.16. Distributions of person INFIT (left) and OUTFIT (right).

be adversely affected. In addition, fit analyses may be adversely affected by smaller sample sizes. For example, fit statistics will have less power, and the creation of empirical/predicted ORFs for fit analysis will be problematic with the small sample sizes produced by the above sample size ratios. As mentioned in previous chapters, sample size requirements for the use of ancillary methods is also a consideration. For instance, if factor analysis is used in dimensionality assessment, then its rules of thumb would be a factor in determining the calibration sample size.

Because of the interaction of the distribution of the respondents across the response categories, as well as across the items, it is difficult to arrive at a hard-and-fast guideline that would be applicable in all situations. However, this is of little consolation to the practitioner. Therefore, for guidance we provide very rough guidelines. Assuming MMLE, a symmetric θ distribution, and the respondents distribute themselves across the response categories in reasonable numbers, we suggest that the minimum sample size be, say 250, in order to address the issues raised above (e.g., fit analysis, facilitating dimensionality assessment, ensuring there are respondents in each category). Moreover, it may be anticipated that there is a sample size, say 1200, at which one reaches, practically speaking, a point of diminishing returns in terms of improvement in estimation accuracy, all other things being equal. (This should not be interpreted as an upper bound.) If one adopts a sample size ratio for determining sample size (e.g., 2 persons per parameter estimated), then it is probably more useful closer to the lower bound of 250 than to the 1200 value (i.e., when the sample size is large, then the sample size ratio becomes less important). These suggestions are tempered by the purpose of the administration (e.g., survey, establishing norms, equating, item pool development/maintenance), the estimation approach, the application's characteristics (e.g., the distribution and range of transition locations, instrument length, latent distribution shape), ancillary technique sample size requirements, and the amount of missing data. As previously mentioned, sample size guidelines should not be interpreted as hard-and-fast rules.

Information for the PC and RS Models

With polytomous models it is possible to determine the amount of information provided by each response category. The sum of these *option information functions* or *category information functions*, $I_{x_j}(\theta)$, across the graded categories (or category scores) is the item information (Samejima, 1969)

$$I_j(\theta) = \sum_{x_j=0}^{m_j} I_{x_j}(\theta)p_{x_j} = \sum_{x_j=0}^{m_j} \frac{(p'_{x_j})^2}{p_{x_j}}, \quad (7.4)$$

where p_{x_j} is the probability of obtaining x_j conditional on θ , and p'_{x_j} is the first derivative of p_{x_j} .²³ The term p_{x_j} may be the PC model or the RS model. For dichotomous data, Equation 7.4 simplifies to the item information formula presented in Chapter 2 (Samejima, 1969). Samejima (1969) shows that there is an increase in item information if a response category is added between two adjacent categories. In short, one obtains greater

item information when treating an item in a polytomous fashion than in a dichotomous fashion. As seen in previous chapters (e.g., Chapter 2), the sum of the item information functions yields the instrument's total information

$$I(\theta) = \sum_{j=1}^L I_j(\theta). \quad (7.5)$$

For the PC model, item information is

$$I_j(\theta) = \sum_{k=1}^{m_j} k^2 p_{x_j} - \left[\sum_{k=1}^{m_j} kp_{x_j} \right]^2 \quad (7.6)$$

and for the RS model, item information is given by

$$I_j(\theta) = \left[\sum_{x=0}^m xp_{x_j} \right]^2 - \sum_{x=0}^m x^2 p_{x_j}. \quad (7.7)$$

The distribution of item information for the PC and RS models differs from that seen with the dichotomous models.²⁴ For instance, with the 1PL model item information has a fixed maximum value that occurs at δ_j . However, for the PC model, items with the same number of score categories provide the same total amount information, although items that have more score categories yield more information across θ than do items with fewer categories (Dodd & Koch, 1987). Moreover, the maximum item information occurs within the range of transition locations. The distribution of item information is affected by the range of the transition locations, the number of reversals of transition locations, and the distance between the reversed transition locations. In general, for a fixed distance between the first and last transition locations, an item that has more transition locations that are in sequential order, and the greater the distance between transition locations that are out of order (if any), the more peaked is the item information function (Dodd & Koch, 1987).

With respect to the RS model's information functions, the location of the maximum of the item information function is affected by the symmetry of the thresholds about the item location, the number of thresholds, and the range of the thresholds. In general, items with four thresholds produce more total information across the continuum than items with three thresholds. Therefore, only items with the same number of thresholds produce the same total information across the θ continuum (Dodd, 1987). The location of the item information maximum is affected by whether the thresholds are symmetrically or asymmetrically distributed about the item's location and whether there is an odd or even number of thresholds. When there is an odd number of asymmetric thresholds (e.g., three), the location of the maximum information shifts away from the item location in the direction of the dominant sign of the thresholds. With an even number of thresholds (e.g., four), the degree of shift in the location of the item information maximum is directly related to the distance between adjacent thresholds. Specifically, the greater the distance between the two middle thresholds, the greater the shift (Dodd & de Ayala, 1994). A different pattern occurs with symmetric thresholds. With an odd number of thresholds that are symmetric about the item's location, the peak of the item

information occurs at δ_j . For items with four thresholds that are symmetric about the item's location, the distance between the middle two thresholds affects the shape as well as the location of the information function's maximum. If the range between the middle two thresholds is less than 2 logits, then the peak of the information function occurs at the item location. However, if the middle two thresholds are greater than 2 logits apart, then the item information may be asymmetrically bimodal. In general, rating scales with thresholds that span a small θ range produce a more peaked information function than when the thresholds have a wider θ range (Dodd, 1987; Dodd & de Ayala, 1994).

Metric Transformation, PC and RS Models

The principles outlined in previous chapters for metric conversion apply to both the PC and RS models. Specifically, the location parameters, δ_{jh} s (or their estimates), are transformed by $\xi^* = \zeta(\xi) + \kappa$, whereas person location parameters (or their estimates) are transformed by $\theta^* = \zeta(\theta) + \kappa$. The methods presented in Chapter 11 can be used to determine the values of ζ and κ .

If one desires to convert PC or RS model-based person locations (or their estimates) to their corresponding expected trait scores and the lowest and highest response categories are represented as 0 and m , respectively, then the expected trait score is

$$T = \sum_{j=1}^L \left[\sum_{k=0}^{m_j} kp_{x_j} \right]. \quad (7.8)$$

The range of T is 0 to $\sum_{j=1}^L m_j$. When Equation 7.8 is applied to the RS model where 1 indicates the lowest response category, then the limits of summation for the inside-the-brackets summation would be 1 and $m + 1$; the range of T is L to $L(m + 1)$.

Summary

The partial credit model is applicable to situations in which one has ordered response data. For example, ordered polytomous data arise in the assignment of partial credit or Likert response data. As such, the PC model is applicable to noncognitive as well as proficiency assessment situations. With the PC model, the ordered responses to an item are referred to as *category scores*. For instance, in a situation where partial credit is assigned, the category scores could be 0, 1, and 2 for “no credit,” “partial credit,” and “full credit,” respectively. With the PC model, items may vary from one another in terms of the number of category scores. Each of these category scores has an associated option response function that describes the probability of obtaining the category score as a function of θ . The intersection of adjacent ORFs (e.g., the ORFs for scores of 0 and 1) occurs at the item's transition locations (δ_{jh} s). The number of transition locations is always one less than the number of category scores. These transition locations do not have to be ordered in terms of magnitude. For example, for an item from a proficiency test, the transition

from no credit to receiving partial credit may be more difficult than the transition from partial to full credit.

A special case of the PC model is the rating scale model. With the RS model, responses are assumed to represent a series of ordered categories (e.g., strongly disagree, disagree, agree, strongly agree). The adjacent response categories are separated by a series of thresholds. In contrast to those of the PC model, these thresholds are constant across items. Therefore, the number of response categories is the same for all items modeled by the rating scale model. Although the thresholds have the same values for all items, their locations on the continuum may vary across the items.

The rating scale model and the partial credit model are members of the Rasch family of models. As such, both the partial credit and rating scale models simplify to the Rasch model when one has two response categories. Moreover, because both models are extensions of the simple Rasch model, they assume that all items are equally effective in discriminating among respondents and neither model addresses the possibility of examinees guessing on items.

In Chapter 8 we present two models that relax the requirement of a common item discrimination. Although the two models differ from one another in how they conceptualize the response process, they both allow item discrimination to vary across items and have item location parameters associated with the response categories. For comparative purposes we apply one model, the generalized partial credit model, to the same example data used with the PC model. The second model, the graded response model, is used with the Attitudes Toward Condoms Scale that we analyzed using the RS model. In contrast to the use of EAP and MLE for person location estimation seen in the previous chapters, in Chapter 8 a third approach for person location estimation, *maximum a posteriori* (MAP), is introduced.

Notes

1. The literature sometimes refers to m_j as the number of “steps” required to correctly answer an item. The term *operation* is used for *step* in the following because the “step” terminology may invite a “sequential steps” interpretation, which contradicts that the partial credit model does not model sequential steps (Andrich, 2015; Masters, 1988; Tutz, 1990).
2. For a two-category item (i.e., $x_j = \{0, 1\}$). Given that

$$\sum_{j=0}^0 (\theta - \delta_{jh}) \equiv 0,$$

then the probability of responding in the highest category (i.e., $x_j = 1$) is

$$p(x_j | \theta, \delta_{jh}) = \frac{\exp \left[\sum_{h=0}^{x_j} (\theta - \delta_{jh}) \right]}{\sum_{k=0}^{m_j} \exp \left[\sum_{h=0}^k (\theta - \delta_{jh}) \right]} = \frac{e^{0+(\theta-\delta_1)}}{\exp \left[\sum_{j=0}^0 (\theta - \delta_{jh}) \right] + \exp \left[\sum_{h=0}^1 (\theta - \delta_{j1}) \right]}$$

$$= \frac{e^{0+(\theta-\delta_1)}}{e^0 + e^{0+(\theta-\delta_1)}} = \frac{e^{(\theta-\delta_1)}}{1 + e^{\theta-\delta_1}}. \quad (7.9)$$

Because when there are two categories there is only one transition location from the lowest category ($x_j = 0$) to the highest category ($x_j = 1$), the subscript h on δ_{jh} may be omitted. The model in Equation 7.9 may be recognized as the Rasch model and indicates that the PC model subsumes the dichotomous Rasch model.

3. To obtain the ORFs presented in Figure 7.2, we calculate the probability of responding in each category as a function of θ . As an example of the relevant calculations, assume that a person is located at 0.0 (i.e., $\theta = 0.0$) and the transition location points are $\delta_{j1} = -1$ and $\delta_{j2} = 1$ for an item with $m = 2$; $x_j = \{0, 1, 2\}$. Therefore, the probability of this individual obtaining a category score of 0 is

$$\begin{aligned} p(x_{j0} = 0 | \theta, \delta_{jh}) &= \frac{\exp \left[\sum_{h=0}^{x_j} (\theta - \delta_{jh}) \right]}{\sum_{k=0}^{m_j} \exp \left[\sum_{h=0}^k (\theta - \delta_{jh}) \right]} = \frac{e^0}{e^0 + e^{0+(0-(-1))} + e^{0+(0-(-1))+(0-1)}} \\ &= \frac{1}{4.7183} = 0.2119. \end{aligned}$$

(The numerator is e^0 because $\sum_{h=0}^0 (\theta - \delta_{jh}) \equiv 0$.)

The probability of this person obtaining a category score of 1 is

$$\begin{aligned} p(x_{j1} = 1 | \theta, \delta_{jh}) &= \frac{\exp \left[\sum_{h=0}^{x_j} (\theta - \delta_{jh}) \right]}{\sum_{k=0}^{m_j} \exp \left[\sum_{h=0}^k (\theta - \delta_{jh}) \right]} = \frac{e^{0+(0-(-1))}}{e^0 + e^{0+(0-(-1))} + e^{0+(0-(-1))+(0-1)}} \\ &= \frac{2.7183}{4.7183} = 0.5761. \end{aligned}$$

For category score 2 we have

$$\begin{aligned} p(x_{j2} = 2 | \theta, \delta_{jh}) &= \frac{\exp \left[\sum_{h=0}^{x_j} (\theta - \delta_{jh}) \right]}{\sum_{k=0}^{m_j} \exp \left[\sum_{h=0}^k (\theta - \delta_{jh}) \right]} = \frac{e^{0+(0-(-1))+(0-1)}}{e^0 + e^{0+(0-(-1))} + e^{0+(0-(-1))+(0-1)}} \\ &= \frac{1}{4.7183} = 0.2119. \end{aligned}$$

The sum of these probabilities across these category scores conditional on $\theta = 0.0$ is $0.2119 + 0.5761 + 0.2119 = 1.0$.

4. In general, it is not possible to estimate the parameters for a category that does not have any observations. However, by reparameterizing the PC model, Wilson and Masters (1993) provide a solution for estimating item parameters for these categories with zero frequencies (i.e., “null” categories).
5. flexMIRT shares some characteristics with MULTILOG and has similar estimation capabilities as IRTPRO. However, IRTPRO has graphical capabilities as well as greater flexibility in data importation and its manipulation than flexMIRT.
6. Whenever one executes flexMIRT and the expected output is not obtained, the user should look for misspellings of commands because flexMIRT does not do as extensive error checking of commands as do other commercially available programs. For example, with Version 3.5.1.23689, misspellings of commands (e.g., M3 for M2, FisherINF for FisherINF, GIF in lieu of GOF, sawSCO for saveSCO) do not generate an error message of “not recognized command.” Rather, the corresponding requested output is not produced. That said, misspellings of command options (e.g., SCORE = EPA) will generate an error message.
7. One can estimate a constant α across items that may not be equal to 1. The relationship between this constant α PC model and the Masters (Rasch) PC model is analogous to the relationship between the 1PL and Rasch models. That is, mathematically, the constant α PC model and the Masters PC model are equivalent: The values from one model can be converted into the other by appropriate rescaling (cf. Chapter 4). To estimate a common discrimination across our eight items, we impose the constraint EQUAL (i1-i8), slope:

```

:
<Options>
  Mode = Calibration;
  GOF = Extended;
:
  SCORE= SSC;           // calculate EAP person location estimates &
  produce conversion table
  savesCO= Yes;         // calculate & save person location estimates
:
<Groups>
  %OnlyGroup%
  File = "ALIKE.DAT";   // space delimited file
  Varnames = i1-i8;
  N = 2942;
  Ncats(i1-i8) = 3;     // 3 valid responses
  Model(i1-i8) = GPC(3);

<Constraints>
  EQUAL (i1-i8), slope; // constrain the GPC slopes to be equal

```

The corresponding output is:

GPC Items for Group 1: OnlyGroup													
Item	Label	P#	a	s.e.	b	s.e.	d	1	d	2	s.e.	d	s.e.
1	i1	17	0.838	0.019	-2.005	0.062	0	-0.454	0.080	0.454	0.080		
2	i2	17	0.838	0.019	-1.050	0.042	0	-0.539	0.065	0.539	0.065		
3	i3	17	0.838	0.019	-0.760	0.041	0	0.482	0.049	-0.482	0.049		
4	i4	17	0.838	0.019	0.191	0.037	0	0.524	0.047	-0.524	0.047		

```
5      i5 17 0.838 0.019 0.220 0.034      0 -1.674 0.090 1.674 0.090
6      i6 17 0.838 0.019 1.450 0.060      0 1.665 0.065 -1.665 0.065
7      i7 17 0.838 0.019 0.948 0.040      0 -0.961 0.073 0.961 0.070
8      i8 17 0.838 0.019 1.573 0.053      0 0.058 0.062 -0.058 0.062
:
```

Marginal reliability for response pattern scores: 0.70

Statistics based on the loglikelihood of the fitted model:

-2loglikelihood: 39370.52

Akaike Information Criterion (AIC): 39404.52

Bayesian Information Criterion (BIC): 39506.30

Full-information fit statistics of the fitted model:

Degrees

G2 of freedom	Probability	F0hat	RMSEA
3148.60	886	0.0001	1.0702
			0.03

The table is too sparse to compute the Pearson X2 statistic.

Even though G2 is shown, it should be interpreted with caution.

Limited-information fit statistics of the fitted model:

Degrees

M2 of freedom	Probability	F0hat	RMSEA
238.12	19	0.0001	0.0809
			0.06

Note: M2 is based on ordinal 1st-order and 2nd-order subtables.

Note: Model-based weight matrix is used.

As can be seen, $\hat{\alpha} = 0.838$ for all items.

In the <Options> section, we specify SSC (Summed Score Conversions) for person location estimation. The SSC will produce individual EAP $\hat{\theta}$ s and a conversion table to transform observed scores X (a.k.a., summed scored) to $\hat{\theta}$ s. This option's output is found in an appropriately labeled tab and in the 's-ssc' file. The relevant part of the '-ssc' file is

:

Summed Score to Scale Score Conversion Table:

Summed

Score	EAP	SD	P
0.00	-2.113	0.638	0.0070398
1.00	-1.788	0.609	0.0154718
2.00	-1.489	0.586	0.0312744
3.00	-1.210	0.569	0.0489704
4.00	-0.945	0.556	0.0683248
5.00	-0.691	0.545	0.0861483
6.00	-0.446	0.538	0.0999299
7.00	-0.206	0.532	0.1063927
8.00	0.030	0.530	0.1060145
9.00	0.265	0.530	0.1008855
10.00	0.502	0.533	0.0915813
11.00	0.743	0.541	0.0786927
12.00	0.993	0.552	0.0628094
13.00	1.255	0.568	0.0472483
14.00	1.535	0.589	0.0296400
15.00	1.838	0.616	0.0157173
16.00	2.173	0.649	0.0038588

Marginal reliability of the scaled scores for summed scores =
0.69963

PC calibration, ALIKE data-equal slope
Item & person parameter estimates, conversion table - single run

Group Parameter Estimates:

Group	Label	mu	s2	sd
1	OnlyGroup	0.000	1.000	1.000

:

Therefore, an individual who incorrectly answered all items ($X = 0$) would have a $\hat{\theta} = -2.113$, a person who received partial credit on one item ($X = 1$) would have a $\hat{\theta} = -1.788$, and so on up to a person who correctly answered all items ($X = 16$) who would have a $\hat{\theta} = 2.173$. This table has many uses. For example, assume we administered our Alike instrument to different individuals. Once we determined their observed score, the table could be used to determine their $\hat{\theta}$ s without having to recalibrate and score their responses.

8. These matrices are so named because the nominal response model uses a slope-intercept formulation (see Chapter 2). Therefore, T_a is the transformation matrix for the slopes (i.e., α s), and T_c is the transformation for the intercepts (i.e., γ); “c” is used because Bock (1972) labeled the intercepts c. These matrices are used to implement identification constraints. See Chapter 9 for the nominal response model as well as the corresponding MULTILOG output on the website for more information. For greater detailed information, see Thissen and Steinberg (1986), Muraki (1992), and Thissen and Cai (2016).

9. With respect to conditional independence, one could use Q_3 or Q_3^P . Recall from Chapter 6 that Q_3 is the correlation between the residuals for a pair of items. In the case of the PC model, the residual for an item is the difference between an individual's observed category score and the individual's expected category score on the item. As such, after fitting the model, the Pearson correlation coefficient is used to examine the linear relationship between pairs of item residuals. In the current situation, the observed category score is a 0, 1, or 2, and the expected response is given by the weighted sum of the response category probabilities according to the PC model (i.e., $d_{ij} = x_{ij} - \varepsilon(x_j | \hat{\theta}_i)$ and $d_{iz} = x_{iz} - \varepsilon(x_z | \hat{\theta}_i)$

$$\mathcal{E}(x_j | \hat{\theta}_i) = \sum_{k=0}^m kp(x_j | \hat{\theta}_i, \delta_{jh}).$$

Symbolically, the residual for person i for item j is $d_{ij} = x_{ij} - \varepsilon(x_j | \hat{\theta}_i)$, and for item z it is $d_{iz} = x_{iz} - \varepsilon(x_z | \hat{\theta}_i)$. Q_3 is the correlation between d_{ij} and d_{iz} across persons.

10. To perform a partial credit calibration in which item discrimination is estimated to be constant but not equal to 1.0, one imposes a constraint on the generalized partial credit model (Chapter 8) to estimate a common α . We use the `mirt.model` function with the `constrain` argument to impose a common estimated α across all eight items. Table 7.12 shows that our α is 0.837. Comparing our Rasch PC model (Table 7.3) and PC model (Table 7.12) item parameter estimates shows the stretch-

TABLE 7.12. mirt Session for the PC Model Calibration of the Alike Data

```

> library(mirt)
  Loading required package: stats4
  Loading required package: lattice

> ConstDiscr=mirt.model('Theta=1-8
+ CONSTRAIN = (1-8,a1)')
> pcm=mirt(alikedata,model=ConstDiscr,itemtype="gpcm",SE=T,SE.type='Fisher')
Iteration: 13, Log-Lik: -19685.262, Max-Change: 0.00010

  Calculating information matrix...
> pcm
  mirt(data = alikedata, model = ConstDiscr, itemtype = "gpcm",
        SE = T, SE.type = "Fisher")

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 13 EM iterations.
mirt version: 1.30
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Information matrix estimated with method: Fisher
Condition number of information matrix = 102.5199
Second-order test: model is a possible local maximum

Log-likelihood = -19685.26
Estimated parameters: 24
AIC = 39404.52; AICc = 39404.73
BIC = 39506.3; SABIC = 39452.29
G2 (6543) = 3148.61, p = 1
RMSEA = 0, CFI = NaN, TLI = NaN>

> M2(pcm,CI=0.95)
      M2 df p      RMSEA   RMSEA_2.5 RMSEA_97.5      SRMSR
  stats 251.1841 19 0 0.06446025 0.05612192 0.0730222 0.06326206
          TLI      CFI
  stats 0.8682644 0.8748511

> coef(pcm,simplify=TRUE,IRTpars=TRUE)
  $items
    a      b1      b2
  I1 0.837 -1.549 -2.461
  I2 0.837 -0.511 -1.590
  I3 0.837 -1.243 -0.278
  I4 0.837 -0.333  0.716
  I5 0.837  1.895 -1.454
  I6 0.837 -0.215  3.117
  I7 0.837  1.911 -0.013
  I8 0.837  1.516  1.631

  $means
  Theta
    0

  $cov
    Theta
    Theta     1

> # obtain person estimates via fscores & display first 6 cases
> people_pcm=fscores(pcm,method="EAP",full.scores=T,full.scores.SE=T)

```

(continued)

TABLE 7.12. (*continued*)

```

> head(people_pcm,6)
      F1      SE_F1
[1,] -1.0531797 0.5489122
[2,] -1.0205381 0.5821920
[3,]  0.1206752 0.5443668
[4,]  0.4707603 0.5467836
[5,]  0.1962325 0.5404228
[6,]  1.8447204 0.6109054

> tail(people_pcm,4)
      F1      SE_F1
[2939,] -0.6404707 0.5674401
[2940,]  0.4585353 0.5101751
[2941,]  1.6259006 0.6204489
[2942,]  1.1371648 0.5632447

> marginal_rxx(pcm)
[1] 0.6979139

> # empirical reliability/marginal reliability
> fscores(pcm,method="EAP",full.scores=T,full.scores.SE=T,returnER=T)
      F1
0.6998303

> mean(people_pcm[,1])    # average person estimate
[1] 0.000132666

> sd(people_pcm[,1])      # SD person estimate
[1] 0.8368019
>
> # obtain person fit info via personfit & display first 6 cases
> head((people_pcmFit=personfit(pcm,method="EAP")), 6)
      outfit   z.outfit   infit   z.infit      Zh
1 0.5500471 -0.5632001 0.6281346 -0.6789048  0.05073361
2 0.5780073 -0.5033801 0.7411355 -0.3920996  1.24169477
3 1.1219227  0.3959105 1.2475491  0.6728296  0.29414575
4 0.5591764 -0.5969808 0.6703927 -0.7689099  1.11324667
5 0.5489979 -0.7109502 0.6008100 -0.9836368  1.31088972
6 0.8988318  0.3077826 1.0843035  0.3412067 -0.76508962

> tail(people_pcmFit,4)
      outfit   z.outfit   infit   z.infit      Zh
2939 0.6979621 -0.34611401 0.7899819 -0.28885391  0.9911905
2940 0.3291015 -1.21324465 0.3726835 -1.89902559  0.5234301
2941 0.7409176  0.06926554 0.9396266  0.07469734  0.5483396
2942 1.5934690  0.85727922 1.1199349  0.39827397 -1.6100062

> # -----
> # model comparisons -----
> # -----
```

> anova(pcm,raschpcm)

```

Model 1: mirt(data = alikedata, model = ConstDiscr, itemtype = "gpcm",
  SE = T, SE.type = "Fisher")
Model 2: mirt(data = alikedata, model = 1, itemtype = "Rasch", SE = T,
  SE.type = "Fisher")

      AIC      AICc      SABIC       HQ       BIC      logLik      X2      df      p
1 39404.53 39404.73 39452.29 39441.17 39506.3 -19685.26      NaN      NaN      NaN
2 39404.52 39404.73 39452.28 39441.17 39506.3 -19685.26  0.001      0      0

```

ing/contracting of the metric due to using $\alpha = 1.0$ vs. $\alpha = 0.837$. Because the difference between the Rasch PC model and the PC model is simply one of metric, the information criteria reflect the equivalence (`anova(pcm, raschpcm)`); any differences in the AIC, BIC, etc. are due to rounding error.

11. X^2 is calculated using the residuals based on the item parameter estimates for predicting the expected frequencies. To obtain Cramér's V , it appears that `mirt` takes the absolute value of X^2 , transforms X^2 to V (e.g., $V = \sqrt{X^2/(N(m_j - 1))}$), and carries forward the sign to V . As such, to determine if we have a large V , we look at its absolute value; $0 \leq |V| \leq 1$.
12. With models that do not have a common α , it would not make sense to do this because the probability of a response is not solely a function of the item location but also involves α .
13. An alternative approach is to find a case with the same X as the case of interest but with `INFIT/OUTFIT` values close to zero. We would then compare the responses between the two cases to find the discrepancies. For example, a comparison case ($X = 13$, `OUTFIT = 0.18`, `INFIT = 0.19`) has a response pattern (22222111). This response pattern is consistent with what one would expect for a respondent that is located around 1.05 (i.e., full credit on items located at and below 1.05 and progressively less credit on items located above 1.05).
14. Although the model requires that each item have the same number of categories, it is also important to note that, conceptually, the items should have response scales that are *functionally* equivalent. For example, consider an instrument that uses a four-category response scale. For some items, this 4-point scale is a Likert scale with the labels of 0 = "strongly disagree," 1 = "disagree," 2 = "agree," 3 = "strongly agree," whereas for other items a 0 represents "never," a 1 represents "sometimes," a 2 reflects "often," and a 3 indicates "always." These two response scales differ functionally from one another. Therefore, it may be argued that whenever the response scale changes across an item set, one might be measuring a different construct or least a different facet of the construct. In these situations, dimensionality analysis may reveal a multidimensional situation. If each dimension consists of items that share a common response scale, then the RS model could be applied to each dimension.
15. Muraki (1992) presents a generalized rating scale model that utilizes a discrimination parameter. This model is

$$p(x_{jk} | \theta, \delta_j, \tau) = \frac{\exp\left[\sum_{k=1}^h \alpha_j(\theta - (\delta_j + \tau_k))\right]}{\sum_{c=1}^m \exp\left[\sum_{k=1}^c (\theta - (\delta_j + \tau_k))\right]},$$

where m is the number of response categories. This model may be estimated using `mirt`, `flexMIRT`, `MULTILOG`, `IRTPRO`, and `PARSCALE`.

16. To obtain the ORFs presented in Figures 7.12 and 7.13, we calculate the probability

of responding in each category as a function of θ . To demonstrate this, assume that an individual is located at $\theta = 0.0$, our item is located at -0.98 ($\delta_1 = -0.98$), and the thresholds have values of $\tau_1 = -0.30$, $\tau_2 = -0.02$, and $\tau_3 = 0.32$.

For convenience of presentation the category coefficient values (κ_{xj} s) are determined first, followed by calculating the denominator of the RS model. From the above we have that

$$\kappa_{xj} = -\sum_{h=1}^{x_j} \tau_h$$

and $\kappa_{0j} = 0$ when $x = 0$. For $x = 1$ this means that

$$\kappa_{1j} = -(0 + (-0.30)) = 0.30.$$

For $x = 2$ we have

$$\kappa_{2j} = -(0 + (-0.30) + (-0.02)) = 0.32$$

and for $x = 3$ we have

$$\kappa_{3j} = -(0 + (-0.30) + (-0.02) + 0.32) = 0$$

The denominator, Υ , is the sum of the four possible numerators

$$\begin{aligned} \Upsilon &= e^{[0 + 0(0 - (-0.98))]} + e^{[0.30 + 1(0 - (-0.98))]} + \\ &\quad e^{[0.32 + 2(0 - (-0.98))]} + e^{[0 + 3(0 - (-0.98))]} \\ &= e^0 + e^{1.28} + e^{2.28} + e^{2.94} = 33.2892. \end{aligned}$$

Therefore, the probability of responding in the “strongly disagree” category for an individual located at $\theta = 0.0$ is

$$p(x_j = 0 | \theta = 0.0, \delta_1, \tau_0) = \frac{\exp[\kappa_{xj} + x_j(\theta - \delta_j)]}{\sum_{k=0}^m \exp[\kappa_k + k(\theta - \delta_j)]} = \frac{\exp[0 + 0(0 - (-0.98))]}{\Upsilon} = \frac{1}{33.2892} = 0.0300.$$

The probability of this individual responding in the “disagree” category ($x = 1$) is

$$p(x_j = 1 | \theta = 0.0, \delta_1, \tau_1) = \frac{\exp[0.30 + 1(0 - (-0.98))]}{\Upsilon} = \frac{3.5966}{33.2892} = 0.1080.$$

For the “agree” category ($x = 2$) the probability is

$$p(x_j = 2 | \theta = 0.0, \delta_1, \tau_2) = \frac{\exp[0.32 + 2(0 - (-0.98))]}{\Upsilon} = \frac{9.7767}{33.2892} = 0.2937$$

and for the “strongly agree” category ($x = 3$) we have

$$p(x_j = 3 | \theta = 0.0, \delta_1, \tau_3) = \frac{\exp[0 + 3(0 - (-0.98))]}{r} = \frac{18.9158}{33.2892} = 0.5682$$

The sum of these probabilities across these category scores conditional on $\theta = 0.0$ is $0.0300 + 0.1080 + 0.2937 + 0.5682 = 1.0$.

17. Our PC and RS models can be extended in situations where one has three-mode data. For example, in a clinical setting, a nurse supervisor will rate the performance of nurses (e.g., on communication with patients, medication administration, skill utilization and development) by using a rating scale across a set of items. Our three modes or facets are supervisor, supervisee, and items. The Many-Facet Rasch Model (MFRM; Linacre, 1994) or Facet model extends the Rasch model to have more than two facets (i.e., facet 1: items and facet 2: individuals).

The MFRM provides an estimate of the location of each element of each facet (e.g., each individual, each rater, and each item is characterized by a parameter). Following Linacre (1994), the MFRM specifies the probability that an individual i rated by judge/rater r will receive a rating in a particular category (k) on a specific item j is

$$\ln\left(\frac{p_{ijr(k)}}{p_{ijr(k-1)}}\right) = \theta_i - \delta_j - H_r - \tau_{jk}. \quad (\text{logistic odds ratio representation})$$

Alternatively, we can represent the model as

$$p_{ijr(k)} = \frac{\exp\left[\sum_{k=0}^x (\theta_i - \delta_j - H_r - \tau_{jk})\right]}{\sum_{q=1}^m \exp\left[\sum_{k=0}^x (\theta_i - \delta_j - H_r - \tau_{jk})\right]}, \quad (\text{probability representation})$$

where

$p_{ijr(k)}$ the probability of individual i being awarded on item j by judge r
a rating of k

$p_{ijr(k-1)}$ the probability of individual i being awarded on item j by judge r
a rating of $k - 1$

θ_i individual i 's location on the construct of interest (facet 1)

δ_j item j 's location on the construct of interest (facet 2)

H_r (capital eta) severity of judge/rater r (facet 3)

τ_{jk} the location (e.g., difficulty to endorse) of the step up from category $k - 1$
to category k on item j . (With a common rating scale across items τ_{jk}
would be replaced by τ_k)

For individuals and items, we have location estimates as detailed with the Rasch model. The model shows that the judges/raters are also located on the same continuum as items and individuals. The judge values (H_r) range from severe (continuum's positive end) to lenient (continuum's negative values). The rating scale is zero-based with m thresholds such that we have ratings $x = 0, 1, \dots, m$. As such, there are $m + 1$ rating categories (e.g., $m = 3$). The model can be extended to a fourth facet. The

Facets program and the R package TAM will provide parameter estimates. Verhelst and Verstralen (2001) present a model for multiple raters.

18. Some authors use the coding $0, \dots, 3$ rather than $1, \dots, 4$ to be consistent with the model's presentation. For example, see Andrich (1978c) and Dodd (1990).
19. Our flexMIRT calibration requires that we recode our 1-based responses (i.e., $1 = \text{"strongly disagree,"}$ $2 = \text{"disagree more than I agree,"} \dots, 4 = \text{"strongly agree"}$) to be 0-based (i.e., $0 = \text{"strongly disagree,"}$ $1 = \text{"disagree more than I agree,"} \dots, 3 = \text{"strongly agree"}$) by using the Code command. Similar to our PC model calibration, for the RS model we impose constraints on the nominal categories model by first fixing the slope for all six items (FIX (i1-i6), slope) and specifying its value to be 1.0 (VALUE . . . 1.0). We also fix all the scoring function contrasts (FIX . . . ScoringFn) and specify the $(m - 1)$ intercepts to be equal across items (EQUAL . . . Intercept . . .).

```

<Project>
  Title = "Rasch RS calibration, Attitudes Towards Condoms Scale ";
  Description = "Item & person parameter estimates - single run ";

<Options>
  Mode = Calibration;
  GOF = Complete;
  :

<Groups>
  %OnlyGroup%
  File = "C:\condomsSpcDlmt.dat";    // space delimited file; no id
  field
  Varnames = i1-i6;
  N = 3473;
  Code(i1-i6) = (1,2,3,4),(0, 1, 2, 3);      // recode responses to be
  0-based
  Ncats(i1-i6) = 4;    // 4 valid responses
  Model(i1-i6) = NOMINAL(4);

<Constraints>
  FIX (i1-i6), slope;
  VALUE (i1-i6), slope, 1.0; // constrain the slopes to be equal to
  1
  FIX (i1-i6), ScoringFn;
  EQUAL (i1-i6), Intercept(2);
  EQUAL (i1-i6), Intercept(3);

```

The corresponding output is

```

  :
  Number of free parameters:      8
  :
  GPC Items for Group 1: OnlyGroup
  Item  Label  P#   a   s.e.   b   s.e.   d   1   d   2   s.e.   d   3   s.e.   d   4   s.e.
  1      i1    1.000  0.112  0.025  0  -0.687  0.024  0.286  0.029  0.401  0.022
  2      i2    1.000  0.669  0.026  0  -0.687  0.024  0.286  0.029  0.401  0.022
  3      i3    1.000  0.783  0.027  0  -0.687  0.024  0.286  0.029  0.401  0.022

```

```
4      i4    1.000  -0.110  0.025   0  -0.687  0.024  0.286  0.029  0.401  0.022  
5      i5    1.000  -0.072  0.025   0  -0.687  0.024  0.286  0.029  0.401  0.022  
6      i6    1.000  -0.358  0.025   0  -0.687  0.024  0.286  0.029  0.401  0.022  
:  
:
```

Marginal reliability for response pattern scores: 0.79

Statistics based on the loglikelihood of the fitted model:

-2loglikelihood: 49964.80

Akaike Information Criterion (AIC): 49980.80

Bayesian Information Criterion (BIC): 50030.02

The Number of free parameters is 8 because we have 6 $\hat{\delta}_j$ plus two $\hat{\tau}_h$ s (the third τ_h is determined by the sum of the two τ_h s: $\sum \tau_h = 0$; α is not estimated. Our $\hat{\delta}_j$ s are found in the column labeled b (i.e., $\hat{\delta}_1 = 0.112$, $\hat{\delta}_2 = 0.669$, . . . , $\hat{\delta}_6 = -0.358$) with d 2, d 3, and d 4 containing our thresholds (i.e., $\hat{\tau}_1 = -0.687$, $\hat{\tau}_2 = 0.286$, and $\hat{\tau}_3 = 0.401$). Given flexMIRT's parameterization, we apply $\delta_{jh} = \delta_j - \tau_h$ to determine the transition point from one category to the next category for each item. For instance, $\hat{\delta}_{11} = \delta_1 - \tau_1 = 0.112 - (-0.687) = 0.799$, $\hat{\delta}_{12} = \delta_1 - \tau_2 = 0.112 - 0.286 = -0.174$, and $\hat{\delta}_{13} = \delta_1 - \tau_3 = 0.112 - 0.401 = -0.289$. The $\hat{\delta}_j$ s show correlations of 1.0 with those of BIGSTEPS. To estimate a common discrimination across items, one removes the FIX (i1-i6), slope and VALUE (i1-i6), slope, 1.0 lines; this is done in Endnote 20.

20. Our flexMIRT calibration for estimating a common across items is:

```
<Project>  
Title = "RS calibration, est common descr; Attitudes Towards Condoms  
Scale";  
Description = "6 Items RS; common slope";  
  
<Options>  
Mode = Calibration;  
savePRM= Yes;           // save item parameter estimates  
FisherInf=81,4.0;       // 81 equal intervals from -4 to 4 to calcu-  
late info  
saveINF= Yes;           // save information function - items/scale  
SE = Fisher;  
  
<Groups>  
%Group1%  
File = "C:\condomsSpcDlmt.dat";  
Varnames = i1-i6;  
N = 3473;  
Code(i1-i6) = (1,2,3,4),(0,1,2,3); // recode to be 0-based  
Ncats(i1-i6) = 4;  
Model(i1-i6) = Nominal(4);  
  
<Constraints>  
Fix (i1-i6),ScoringFn;  
Equal (i1-i6),Slope;
```

```
Equal (i1-i6),Intercept(2);
Equal (i1-i6),Intercept(3);
```

The corresponding output is

```
:
Number of free parameters: 9
:

GPC Items for Group 1: Group1
Item Label P# a s.e. b s.e. d 1 d 2 s.e. d 3 s.e. d 4 s.e.
1 i1 7 0.50 0.01 0.20 0.03 0 -1.68 0.07 0.57 0.06 1.12 0.06
2 i2 7 0.50 0.01 1.12 0.04 0 -1.68 0.07 0.57 0.06 1.12 0.06
3 i3 7 0.50 0.01 1.31 0.05 0 -1.68 0.07 0.57 0.06 1.12 0.06
4 i4 7 0.50 0.01 -0.17 0.03 0 -1.68 0.07 0.57 0.06 1.12 0.06
5 i5 7 0.50 0.01 -0.11 0.03 0 -1.68 0.07 0.57 0.06 1.12 0.06
6 i6 7 0.50 0.01 -0.59 0.04 0 -1.68 0.07 0.57 0.06 1.12 0.06
:
Marginal reliability for response pattern scores: 0.63

Statistics based on the loglikelihood of the fitted model:
-2loglikelihood: 49053.64
Akaike Information Criterion (AIC): 49071.64
Bayesian Information Criterion (BIC): 49127.02
```

The Number of free parameters is 9 because we have one $\hat{\alpha}$, 6 $\hat{\delta}_j$ s plus two $\hat{\tau}_h$ s (the third $\hat{\tau}_3$ is determined by the sum of the two $\hat{\tau}_h$ s because $\sum \tau_h = 0$). The common discrimination estimate ($\hat{\alpha}$) is found in the column labeled a (i.e., $\hat{\alpha} = 0.50$). Our location estimates $\hat{\delta}_j$ s are found in the column labeled b (i.e., $\hat{\delta}_1 = 0.20$, $\hat{\delta}_2 = 1.12$, \dots , $\hat{\delta}_6 = -0.58$) with d2, d3, and d4 containing our thresholds (i.e., $\hat{\tau}_1 = -1.68$, $\hat{\tau}_2 = 0.57$, $\hat{\tau}_3 = 1.12$). Given flexMIRT's parameterization, we apply $\delta_{jh} = \hat{\delta}_j - \tau_h$ to determine the transition point from one category to the next category for each item. For instance, $\hat{\delta}_{11} = \delta_1 - \tau_1 = 0.2 - (-1.68) = 1.88$, $\hat{\delta}_{12} = \delta_1 - \tau_2 = 0.2 - 0.57 = -0.37$, and $\hat{\delta}_{13} = \delta_1 - \tau_3 = 0.2 - 1.12 = -0.92$. The $\hat{\delta}_j$ s and $\hat{\delta}_{jh}$ s show correlations of 1.0 with those of BIGSTEPS.

21. INFIT and OUTFIT MNSQs can be transformed into either z (ZSTD) or t distributions. In these cases, the range for these transformed fit statistics is $-\infty$ to ∞ with an expectation of 0. However, the interpretation of ZSTD depends on whether rescaling is done, and if so, the type of scaling. The LOCAL keyword is used to specify rescaling. If LOCAL = N, then no rescaling is performed and ZSTD should be ignored. If LOCAL = Y, then the MNSQs are rescaled to reflect their level of significance in the context of the degree of error in the data. In this case, negative ZSTD values indicate a response string close to a Guttman pattern, and positive values indicate more variation than would be predicted by the Rasch model. If LOCAL = L, then the MNSQs are logarithmically rescaled. For this example TABLE 0.1 shows that LOCAL = N, so the ZSTDs are ignored (see Table 7.6). See Linacre and Wright (2001) for more information on this issue.

22. Our mirt RS model calibration uses the itemtype = "rsm"; our response data may be 1-or 0-based (see Endnote 19). Our R session is:

```
> Raschrsm=mirt(condomsdata,model=1,itemtype="rsm",SE=T)
  Iteration: 27, Log-Lik: -24526.820, Max-Change: 0.00009

  Calculating information matrix...
> coef(Raschrsm,simplify=T,IRTpars=T)
$items
  a1     b1     b2     b3      c
I1  1 0.943 -0.186 -0.461  0.000
I2  1 0.943 -0.186 -0.461 -0.464
I3  1 0.943 -0.186 -0.461 -0.558
I4  1 0.943 -0.186 -0.461  0.186
I5  1 0.943 -0.186 -0.461  0.153
I6  1 0.943 -0.186 -0.461  0.393

$means
F1
  0

$cov
F1
F1 0.252
```

The column labeled *c* contains our item easiness estimates ($\hat{\delta}_j^E$); the argument IRTpars = T does not have an effect with the RS model calibration in mirt version 1.30. Given our Likert response scale with 1 = “strongly disagree” and 4 = “strongly agree” and neutrally written items, an “easy to endorse” scale is conceptually appealing. However, if one wishes to work with a “difficulty to endorse” scale, this easiness scale can be reflected to be a “difficulty to endorse” scale by $\delta_j = -\hat{\delta}_j^E$.

As stated in Chapter 3, Endnote 1, one way of handling the indeterminacy of the metric is the standard item approach. mirt implements this strategy by setting the first item’s location to 0 (i.e., $\hat{\delta}_1^E = 0$).

The columns *b1*, *b2*, and *b3* present nondeviation “offsets” and, as a result do not sum to 0. Although the differences between *b1*, *b2*, and *b3* are constant across items, these are not our thresholds per se. We can obtain our thresholds by creating deviations about the mean location. Thus, we calculate the $\text{avg}(\text{offset}) = (0.943 + (-0.186) + (-0.461))/3 = 0.099$ and subtract it from each offset_h to obtain the thresholds $\hat{\tau}_h = \text{offset}_h - \text{avg}(\text{offset})$. This gives us $\hat{\tau}_1 = -0.943 - 0.099 = 0.844$, $\hat{\tau}_2 = -0.186 - 0.099 = -0.285$, and $\hat{\tau}_3 = -0.461 - 0.099 = -0.560$; $\sum \hat{\tau}_h = 0$. The actual transition locations on the continuum are obtained by $\hat{\delta}_{jh}^E = \hat{\delta}_j^E + \hat{\tau}_h$ for each item and each *h*. Calculating our $\hat{\delta}_{jh}^E$ s for all our items and response categories yields

	$\hat{\delta}_{j1}^E$	$\hat{\delta}_{j2}^E$	$\hat{\delta}_{j3}^E$
1	0.844	-0.285	-0.560
2	0.380	-0.749	-1.024
3	0.286	-0.843	-1.118

4	1.030	-0.099	-0.374
5	0.997	-0.132	-0.407
6	1.237	0.108	-0.167

The correlation between the $\hat{\delta}_{jh}^E$ s and our BIGSTEPS' $\hat{\delta}_{jh}^E$ s are each -1.000, as are the correlations between mirt's $\hat{\delta}_j^E$ and BIGSTEPS's $\hat{\delta}_j^E$ as well as flexMIRT's $\hat{\delta}_j$. In short, although our estimates are on a different metric than those of the other programs, our $\hat{\delta}_{jh}^E$ s and $\hat{\tau}_h$ s are highly linearly related to those of the programs and a linear transformation can be used to align our metrics.

23. When one scores an item in a graded fashion (and assuming the operating characteristic of the graded response is twice-differentiable), then the information function for the graded response is (Samejima, 1969)

$$I_{x_j}(\theta) = \left\{ \frac{\partial^2 \ln p_{x_j}}{\partial \theta^2} \right\} p_{x_j} = \frac{(p'_{x_j})^2}{(p_{x_j})^2} - \frac{p''_{x_j}}{p_{x_j}}, \quad (7.11)$$

where p is the probability of obtaining x_j conditional on θ , p'_{x_j} is the first derivative of p_{x_j} , and p''_{x_j} is the second derivative of p_{x_j} . $I_{x_j}(\theta)$ is the option information function.

The sum of the option information functions across the categories (or category scores) is defined by Samejima (1969) as the item information, $I_j(\theta)$

$$\begin{aligned} I_j(\theta) &= -\mathcal{E} \left\{ \frac{\partial^2 \ln p_{x_j}}{\partial \theta^2} \right\} \\ &= \sum_{x_j=0}^{m_j} I_{x_j}(\theta) p_{x_j} = \sum_{x_j=0}^{m_j} \left\{ \frac{(p'_{x_j})^2}{p_{x_j}} - p''_{x_j} \right\} \\ &= \sum_{x_j=0}^{m_j} \frac{(p'_{x_j})^2}{p_{x_j}} - \sum_{x_j=0}^{m_j} p''_{x_j}. \end{aligned} \quad (7.12)$$

However, because the sum of the second derivatives across m_j graded categories is 0, Equation 7.12 simplifies to Equation 7.4.

24. The standard error for the person location estimate under the PC model is

$$s_e(\hat{\theta}_i) = \sqrt{\frac{1}{\sum_{j=1}^L \left\{ \sum_{k=1}^{m_j} k^2 p_{x_j} - \left[\sum_{k=1}^{m_j} k p_{x_j} \right]^2 \right\}}}, \quad (7.13)$$

and for the RS model it is

$$s_e(\hat{\theta}_i) = \sqrt{\sum_{j=1}^L \left\{ \left[\sum_{x=0}^m x p_{xj} \right]^2 - \sum_{x=0}^m x^2 p_{xj} \right\}}. \quad (7.14)$$

For both Equations 7.13 and 7.14 p_{xj} is conditional on $\hat{\theta}_i$.

8

Non-Rasch Models for Ordered Polytomous Data

In Chapter 7 we discussed ordered polytomous data that are extensions of the dichotomous Rasch model. In this chapter, we present two models that relax the assumption of equal item discrimination and that are applicable to ordered polytomous data.¹ These two models, the generalized partial credit and the graded response models, differ from one another in their conceptualization of the operating characteristic function. We begin with the generalized partial credit model and apply it to the reasoning ability instrument from Chapter 7. In this example, we use *maximum a posteriori* (MAP) for person location estimation. With the presentation of MAP we have presented the three common ability estimation approaches (EAP, MAP, MLE). Subsequently, the graded response model is presented. As part of this presentation, the Attitudes Toward Condoms Scale from Chapter 7 is used to demonstrate applying the graded response model.

The Generalized Partial Credit Model

As mentioned in Chapter 7, the PC model assumes that all items on an instrument have equal discrimination. The *generalized partial credit* (GPC; Muraki, 1992) model relaxes this assumption. Muraki developed his model using, in essence, Masters's (1982) approach. That is, by assuming that the probability of selecting a particular response category over the previous one is governed by a dichotomous model. However, instead of using the dichotomous Rasch model for determining the probability of a response, Muraki used the 2PL model. The result is a model that specifies the probability of providing a response in item j 's k th category, x_{jk}

$$p(x_{jk} | \theta, \alpha_j, \delta_{jk}) = \frac{\exp \left[\sum_{h=1}^{k_j} \alpha_j (\theta - \delta_{jh}) \right]}{\sum_{c=1}^{m_j} \exp \left[\sum_{h=1}^c \alpha_j (\theta - \delta_{jh}) \right]}, \quad (8.1)$$

where θ is the latent trait, α_j is the item discrimination, δ_{jh} is the transition location parameter between the h th category and the $h - 1$ category (i.e., the intersection point of adjacent ORFs), m_j , is the number of categories, and $k = \{1, \dots, m_j\}$. Muraki defines the first boundary location as zero (i.e., $\delta_{j1} \equiv 0$), so there are $m_j - 1$ transition locations. (Note that we use the italicization of “m” to indicate the number of response categories, not the number of transition locations.) For example, a four-response category item would have $m_j = 4$ and three transition location parameters labeled δ_{j2} , δ_{j3} , and δ_{j4} . As is the case with the PC model, the δ_{jh} s do not have to be in sequential order. Some (e.g., Yen, 1993) refer to the model in Equation 8.1 as the two-parameter partial credit (2PPC) model.

As previously mentioned, Masters (1982) shows that the RS model can be obtained from his PC model by decomposing an item's δ_{jh} s into an item location parameter and a threshold component. Similarly, the transition locations in Equation 8.1 may be decomposed into an item location component and a threshold parameter to obtain a formulation of the GPC that is analogous to the RS model but that models items with varying discrimination. By substituting $\delta_{jh} = \delta_j - \tau_h$ into Equation 8.1, we obtain a model that specifies the probability of responding in category k , x_{jk} , on an item j as

$$p(x_{jk} | \theta, \alpha_j, \delta_j, \tau) = \frac{\exp \left[\sum_{h=1}^{k_j} \alpha_j (\theta - \delta_j + \tau_h) \right]}{\sum_{c=1}^{m_j} \exp \left[\sum_{h=1}^c \alpha_j (\theta - \delta_j + \tau_h) \right]}, \quad (8.2)$$

where m_j is the number of response categories, $k = 1, \dots, m$, and $\tau_1 \equiv 0$. As can be seen, this parameterization of the GPC model is similar to the RS model, but without the constraint that item discriminations be equal to 1 for all items. Muraki (1992) considers Equation 8.2 to be the GPC model “unless its rating aspect is specifically emphasized” (p. 165). Therefore, although both Equations 8.1 and 8.2 may be considered the GPC model, because Equation 8.2 is the “rating formulation” of the GPC model, we refer to it as the *generalized rating scale* (GRS) model; also see Muraki (1990). In the following discussion, p_{x_j} is used for $p(x_{jk} | \theta, \alpha, \delta, \tau)$ and $p(x_{jk} | \theta, \alpha_j, \delta_{jk})$.

In Equations 8.1 and 8.2, the discrimination parameter, α_j , “indicates the degree to which categorical responses vary among items as θ changes” (Muraki, 1992, p. 162). Although α_j has a range from $-\infty$ to ∞ , its acceptable range consists of positive values. Muraki interprets the τ_h in Equation 8.2 as the relative difficulty of step h “in comparing other steps within an item” (p. 165); “difficulty” may also be interpreted as the difficulty of endorsing a particular category. Moreover, the τ_h s do not need to be sequentially

ordered for $k = 1, \dots, m$. Parameter estimation for the GRS model may be accomplished using flexMIRT, mirt, or PARSCALE.

To demonstrate the impact of a varying discrimination parameter and transition point locations, we use a series of ORFs. Consistent with the partial credit terminology used in Chapter 7, we use the category scores of 0, 1, and 2 for no credit, partial credit, and full credit, respectively. In terms of the GPC model's parameterization in Equation 8.1, the response categories of 1, 2, and 3 correspond to the figures' labeled category scores of 0, 1, and 2, respectively.

Figures 8.1 to 8.3 show the ORFs for 3 three-category items with identical transition locations, but with different discrimination values; $\alpha_1 = 0.50$, $\alpha_2 = 1.0$, and $\alpha_3 = 2.0$ for Figures 8.1 to 8.3, respectively. Because δ_{j1} is defined as equal to 0, the intersections of the ORFs are labeled (generally speaking) δ_{j2} and δ_{j3} (i.e., $\delta_{j2} = -1.0$ and $\delta_{j3} = 1.0$). Comparing the three figures, one sees that as the item discrimination value increases, the ORFs for the zero and full-credit scores (i.e., category scores of 0 [$x = 0$] and 2 [$x = 2$]) are steeper. As is the case with the PC model, the ORFs for any item always consist of one monotonically nondecreasing ORF (e.g., category 2) and one monotonically nonincreasing ORF (e.g., category 0), with typically one unimodal ORF for each additional response category.

Figures 8.3 and 8.4 contain the ORFs for two items with a common α_j , but with transition parameters farther apart in Figure 8.4 than they are in Figure 8.3. Contrasting these two figures, we see that when the transition parameters are in increas-

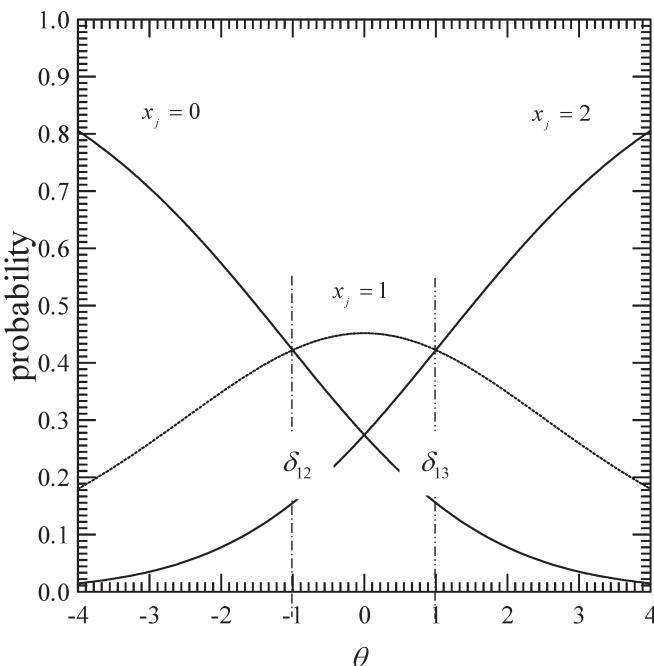


FIGURE 8.1. GPC model ORFs for a three-category item with $\alpha_1 = 0.50$, $\delta_{12} = -1.0$, and $\delta_{13} = 1.0$.

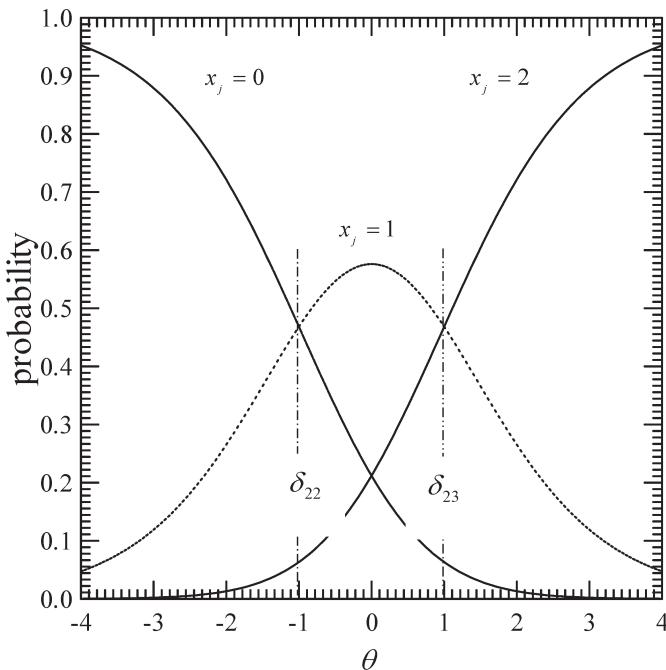


FIGURE 8.2. GPC model ORFs for a three-category item with $\alpha_2 = 1.00$, $\delta_{22} = -1.0$, and $\delta_{23} = 1.0$.

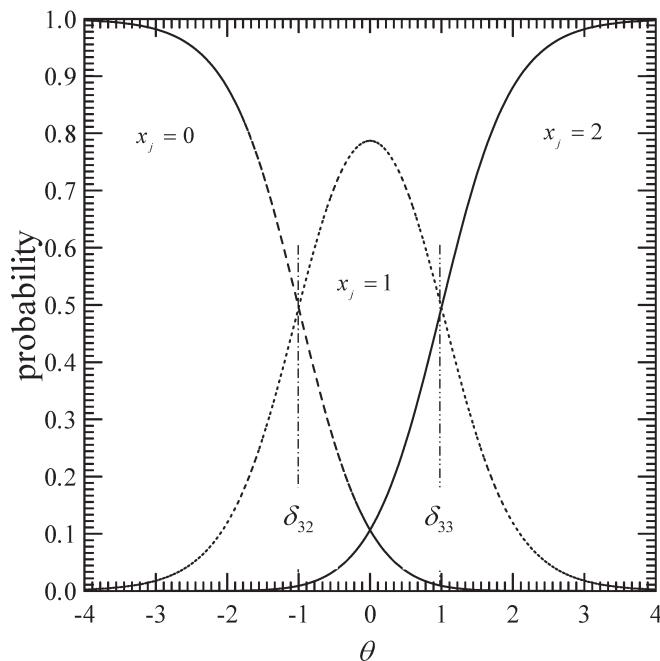


FIGURE 8.3. GPC model ORFs for a three-category item with $\alpha_3 = 2.0$, $\delta_{32} = -1.0$, and $\delta_{33} = 1.0$.

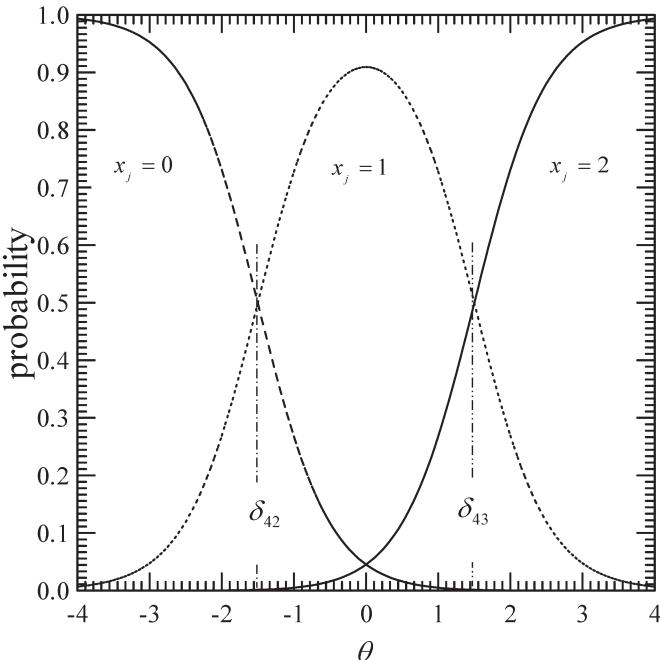


FIGURE 8.4. GPC model ORFs for a three-category item with $\alpha_4 = 2.0$, $\delta_{42} = -1.5$, and $\delta_{43} = 1.5$.

ing order and become farther apart, the probability of obtaining a category score of 1 ($x = 1$) increases throughout the continuum, as compared with when the δ_{jh} s are closer together.

To understand the effect of item discrimination on the ORFs for the nonzero/non-full-credit category scores, we need to attend to whether the transition locations are in increasing order as well as to the distance between transition location parameters. Figures 8.5 and 8.3 present items with the same discrimination, but with the transition parameters reversed. For Figure 8.5 $\delta_{52} > \delta_{53}$, but for Figure 8.3 we have $\delta_{32} < \delta_{33}$. When the transition locations are in sequential order (i.e., $\delta_{j2} < \delta_{j3}$), then as α_j increases, the ORF(s) for the nonzero/non-full-credit response categories become more peaked (cf. Figures 8.1 and 8.3).

When there is a reversal in the transition locations (e.g., $\delta_{52} = 1.0$ and $\delta_{53} = -1.0$, as in Figure 8.5), then the probability of obtaining a category score of 1 is substantially less than that of obtaining a score of either 0 or 2. This is the same pattern observed with reversed transition locations in the PC model and shown in Figure 7.3. In effect, the item represented in Figure 8.5 is almost functioning as a binary item with an item location of 0 (i.e., $(\delta_{52} + \delta_{53})/2$). Consistent with this interpretation is the observation that the ORF for the response category of 1 is not very effective at attracting respondents. In a proficiency assessment situation, such an item indicates that examinees either correctly or incorrectly answered the item with very few individuals receiving partial credit.

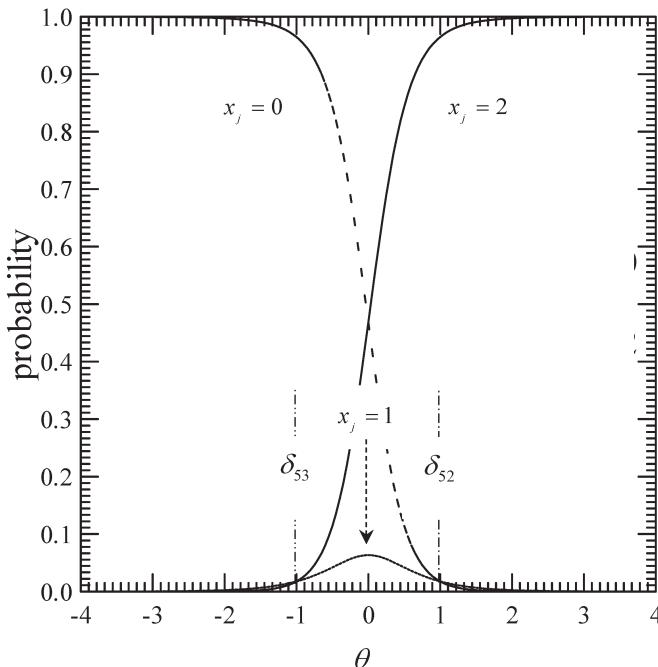


FIGURE 8.5. GPC model ORFs for a three-category item with $\alpha_5 = 2.0$, $\delta_{52} = 1.0$, and $\delta_{53} = -1.0$.

Despite this observation, recall that α_j reflects how well the item discriminates among different values of θ , not how well each response category discriminates. Contrasting Figure 8.5 ($\alpha_5 = 2.0$, $\delta_{52} = 1.0$, $\delta_{53} = -1.0$) with Figure 8.6 ($\alpha_6 = 2.0$, $\delta_{62} = 1.5$, $\delta_{63} = -1.5$) shows that as the transition parameters become farther apart, the ORF for category 1 decreases, while holding α_j fixed.

The GPC model can be shown to be related to other models. For example, with a two-category item (i.e., $m_j = 2$) the GPC simplifies to the 2PL model. Moreover, as is the case with the PC model, the GPC is a special case of Bock's nominal response model (see Chapter 9).

Example: Application of the GPC Model to a Reasoning Ability Instrument, MMLE, flexMIRT

There are several programs available for GPC model estimation (e.g., flexMIRT, mirt, TAM, SAS proc irt). As an example, we apply the GPC model to the Alike Reasoning data that is calibrated with the PC model in Chapter 7. For comparison with the PC model results we use flexMIRT. Following the flexMIRT analysis, we use mirt to perform our calibration. (On the author's website the GPC model analysis of these data with PARSCALE may be found.) As was the case with the PC model, the GPC is implemented as a special case of the nominal model. However, in contrast to the

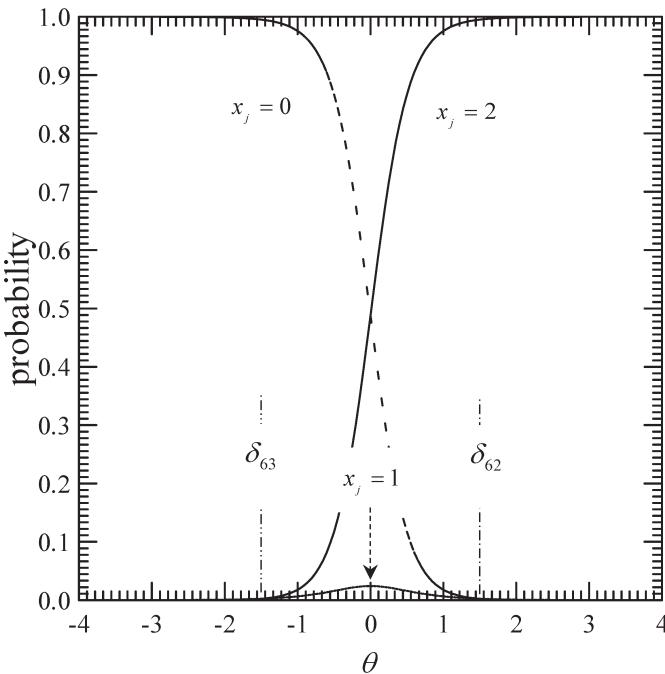


FIGURE 8.6. GPC model ORFs for a three-category item with $\alpha_j = 2.0$, $\delta_{j2} = 1.5$, and $\delta_{j3} = -1.5$.

PC model the GPC model is a primary model. As such, we do not need to specify any constraints, although our output will reflect its nominal model heritage. Unlike our Chapter 7 example, we now separate the item parameter estimation from person parameter estimation.

Assume that we have “cleaned” our data, assessed the tenability of our assumptions as performed above, and wish to fit the GPC model to our data. Table 8.1 shows the command file. As can be seen, the command file is similar to that used for the PC calibration (Table 7.1) except for the removal of the SCORE and saveSCO commands as well as the absence of constraints. (Although we do not specify any constraints, we still need the <Constraints> section header.)

Our abridged output is shown in Table 8.2. Comparing the fit of our GPC and PC models (Chapter 7, Endnote 7; $\alpha = 0.838$) shows that the GPC model’s AIC and BIC are smaller than those of the PC model. Moreover, our GPC model fits significantly better than the PC model with a $\Delta G^2 = 77.25$ on $df = 24 - 17 = 7$. Of course, our next step is to perform model-level and item-level fit analyses. Because the process is the same as presented above, we do not repeat it here.

As can be seen, our discrimination estimates (α) vary from 0.625 to 1.3, with an average discrimination of 0.86375. Our poorest discriminating is item 6 (“In what way is work and play alike?”) with a $\hat{\alpha}_6 = 0.625$. The subsequent columns contain the item location estimate $\hat{\delta}_j$ (labeled b), its standard error (s.e.), and $\hat{\tau}_h$ labeled d (e.g., d 2, d 3)

TABLE 8.1. Command File for the flexMIRT GPC Model Calibration Example—Only Item Parameter Estimation

```

<Project>
  Title = "GPC calibration, ALIKE data, 8 three category items ";
  Description = "Obtain & save item parameter estimates";

<Options>
  Mode = Calibration;
  GOF = Extended;
  M3=ordinal;
  NumDec = 3;
  savePRM= Yes;
  FisherInf= 81, 4.0;
  saveINF= Yes;
  SE= Fisher;

<Groups>
  %OnlyGroup%
  File = "ALIKE.DAT";           // space delimited file
  Varnames = i1-i8;
  N = 2942;
  Ncats(i1-i8) = 3;
  Model(i1-i8) = GPC(3);      // specify GPC model, a primary model

<Constraints>

```

with their corresponding standard errors. We can see that, overall, item 8 is our most difficult item (“In what way is praise and punishment alike?”) and that item 1 is our easiest (“In what way are an orange and a banana alike?”). As is the case with the PC model, we need to subtract our estimated offsets from the item location estimates (i.e., $\hat{\delta}_{jh} = \hat{\delta}_j - \hat{\tau}_h$). Therefore, to obtain $\hat{\delta}_{j1}$ and $\hat{\delta}_{j2}$ we combine item j ’s $\hat{\delta}_j$ with its d2 and d3 values. For example, for item 1 we have $\hat{\delta}_{12} = \hat{\delta}_1 - \hat{\tau}_2 = -1.631 - (-0.089) = -1.542$ and $\hat{\delta}_{13} = \hat{\delta}_1 - \hat{\tau}_3 = -1.631 - (0.089) = -1.720$. All of our estimates are shown in Table 8.3.

Assume that we find the fit of our GPC model to the data is acceptable. As a result, we proceed to estimating our respondents’ locations. Above we have used the MLE and EAP approaches to person location estimation. This time we use *maximum a posteriori* (MAP). (MAP is discussed in Chapter 4.) Recall that like EAP, MAP provides a Bayes estimator that is typically regressed toward the mean of the prior distribution. However, whereas EAP uses the mean of the posterior distribution, MAP uses its mode. As a Bayesian approach, MAP provides person location estimates for all response patterns. Table 8.4 shows the flexMIRT command file for obtaining our MAP $\hat{\theta}$ s, saving the estimates to an external file (saveS = Yes), and the corresponding output. In the <Options> section, we specify the file that contains the item parameter estimates (ReadPRMFile = “GPC-prm.txt”), to estimate only people (Mode = Scoring), and to use the MAP approach (SCORE = MAP); our <Groups> and <Constraints> sections are the same as in Table 8.1. In our output flexMIRT echoes the number of persons, the estimation approach, and the item parameter estimates it read from the ReadPRMFile file. Our $\hat{\theta}$ s are found in the “-sco” file. Our first person required two iterations to achieve convergence and is estimated to be located at -0.941979 with a $s_e(\hat{\theta})$ of 0.506829.

TABLE 8.2. Abridged Output from the flexMIRT GPC Model Calibration of the Alike Data

```
:
GPC calibration, ALIKE data
Obtain & save item parameter estimates
:
Convergence and Numerical Stability
flexMIRT(R) engine status: Normal termination
First-order test: Convergence criteria satisfied
Condition number of information matrix: 38.8058
Second-order test: Solution is a possible local maximum
:
GPC Items for Group 1: OnlyGroup
Item Label P#    a      s.e.    b      s.e.    d 1    d 2      s.e.    d 3      s.e.
  1   i1     1  1.300  0.093 -1.631  0.063    0 -0.089  0.068  0.089  0.068
  2   i2     4  1.059  0.065 -0.935  0.044    0 -0.325  0.069  0.325  0.069
  3   i3     7  0.842  0.050 -0.760  0.048    0  0.480  0.049 -0.480  0.049
  4   i4    10  0.673  0.042  0.214  0.044    0  0.538  0.058 -0.538  0.058
  5   i5    13  0.867  0.052  0.212  0.034    0 -1.603  0.141  1.603  0.141
  6   i6    16  0.625  0.047  1.804  0.127    0  2.058  0.140 -2.058  0.140
  7   i7    19  0.882  0.056  0.922  0.048    0 -0.888  0.105  0.888  0.105
  8   i8    22  0.662  0.048  1.827  0.106    0 -0.047  0.086  0.047  0.086

:
Marginal reliability for response pattern scores: 0.71

Statistics based on the loglikelihood of the fitted model:
          -2loglikelihood: 39293.27
          Akaike Information Criterion (AIC): 39341.27
          Bayesian Information Criterion (BIC): 39484.95

Full-information fit statistics of the fitted model:
          Degrees
          G2 of freedom Probability      F0hat      RMSEA
          3071.35        879       0.0001     1.0440      0.03
The table is too sparse to compute the Pearson X2 statistic.
Even though G2 is shown, it should be interpreted with caution.

Limited-information fit statistics of the fitted model:
The M2 statistics were not requested.
```

Example: Application of the GPC Model to a Reasoning Ability Instrument, MMLE, mirt

Our GPC model calibration of the Alike data (Table 8.5) is a continuation of the PC model R session (see Table 7.12). We specify our itemtype to be gpcm (mirt(alikedata, 1, 'gpcm', . . .)) and assign the results to the output object gpcm. As mentioned above, our discussions of model- and item-level fit as well as person-fit apply here. Our calibration converged in 29 iterations. A comparison of the GPC (Model 2) model-

TABLE 8.3. Item Parameter Estimates from flexMIRT GPC Model Calibration Example

Item	$\hat{\alpha}_i$	$\hat{\delta}_{i2}$	$\hat{\delta}_{i3}$
1	1.300	-1.542	-1.720
2	1.059	-0.610	-1.260
3	0.842	-1.240	-0.280
4	0.673	-0.324	0.752
5	0.867	1.815	-1.391
6	0.625	-0.254	3.862
7	0.882	1.810	0.034
8	0.662	1.874	1.780

level fit information with those of the PC model (Model 1) shows that the GPCM fits comparatively better than the PC model. The GPC model's AIC and BIC are smaller than those of the PC model and our $\Delta G^2 = 77.254$ with $df = 7$.

Our item parameter estimates (`coef(gpcm,)`) and their descriptive statistics are identical (or nearly so) to those of flexMIRT. As mentioned above, with the GPC model, items are free to vary and do vary in their α_j s. Our $\hat{\alpha}_j$ s vary from 0.625 to 1.299, with an average estimated discrimination of 0.864. With our PC calibration (Table 7.12) our common $\hat{\alpha}$ is 0.837.

We use the `mirt's plot` function to obtain the ORFs for all eight items (Figure 8.7). Contrasting the items with one another, we see the impact of the varying $\hat{\alpha}_j$ s and the distance between an item's $\hat{\delta}_{jh}$ s. As an example, item 3's $\hat{\delta}_{3h}$ s are in order, whereas those of item 1's show a reversal. As a consequence, the item 3's partial credit ORF (P2; the unimodal ORF; light gray) shows that individuals located in the neighborhood of -1 are most likely to receive partial credit on the question "In what way are an eye and an ear alike?" whereas above approximately 0 they will most likely receive full credit and below approximately -1.25 they will most likely not receive any credit. In contrast, on item 1 "In what way are an orange and a banana alike?" respondents tend to receive no credit (P1; left monotonically decreasing ORF; medium gray) or full credit (P3; right monotonically increasing ORF; dark gray). In fact, several of our items (1, 2, 5, 7, 8) behave in a dichotomous fashion, even though on each of these items at least 259 individuals were given partial credit (cf. Table 7.3).

To parallel our flexMIRT calibration, we obtain MAP person location estimates (`fscores(gpcm, method = "MAP",)`). We display the first six and last four cases. For example, our first case is estimated to be at -0.942 ($s_e(\hat{\theta}) = 0.507$), our second $\hat{\theta}_2 = -0.767$ ($s_e(\hat{\theta}) = 0.509$), and so on. For pedagogical reasons we also obtain our EAP $\hat{\theta}$ s. As can be seen, our EAP $\hat{\theta}$ s are very similar to our MAP $\hat{\theta}$ s and show a strong linear relation ($r = 1.0$). Their descriptive statistics show similar means and

TABLE 8.4. Command File for the flexMIRT GPC Model Calibration Example—Only Person Parameter Estimation

```

Project>
  Title = "GPC calibration, ALIKE data ";
  Description = "Obtain & save person parameter estimates";

<Options>
  Mode = Scoring;                                // just estimate persons
  GOF = Extended;
  ReadPRMFile= "GPC-prm.txt";                    // read previously estimated item parameter
  SCORE= MAP;                                     // calculate MAP person location estimates
  saveS= Yes;                                     // save person location estimates; abbreviate saveSCO

<Groups>
  %OnlyGroup%
  File = " ALIKE.DAT";
  Varnames = i1-i8;
  N = 2942;
  Ncats(i1-i8) = 3;
  Model(i1-i8) = GPC(3);

<Constraints>

```

The following output is from 'GPC-ssc.txt' file.

```

flexMIRT(R) Engine Version 3.51 (64-bit)
Flexible Multilevel Multidimensional Item Response Modeling and Test Scoring
(C) 2013-2017 Vector Psychometric Group, LLC., Chapel Hill, NC, USA

```

```

GPC calibration, ALIKE data
Obtain & save person parameter estimates

```

Summary of the Data and Dimensions

Missing data code	-9
Number of Items	8
Number of Cases	2942
# Latent Dimensions	1

:

Scoring Control Values

Response pattern MAPs are computed

Miscellaneous Control Values

Output Files

Text results and control parameters: GPC-ssc.txt

Text scale score file: GPC-sco.txt

GPC calibration, ALIKE data

Obtain & save person parameter estimates

GPC Items for Group 1: OnlyGroup

Item	a	b	d	1	d	2	d	3
1	1.30044-1.63139			0-0.08887	0.08887			
2	1.05947-0.93487			0-0.32545	0.32545			
3	0.84174-0.76044			0	0.48034-0.48034			
4	0.67305	0.21441		0	0.53782-0.53782			
5	0.86653	0.21238		0-1.60296	1.60296			
6	0.62543	1.80363		0	2.05824-2.05824			
7	0.88150	0.92247		0-0.88770	0.88770			
8	0.66246	1.82721		0-0.04669	0.04669			

:

Marginal reliability for response pattern scores: 0.71

(continued)

TABLE 8.4. (*continued*)

The following are the MAP $\hat{\theta}$ s found in the saved '-sco.txt' file. The format is Group, record's ordinal position in the data file, the number of iterations to obtain $\hat{\theta}$, $\hat{\theta}$, and $s_e(\hat{\theta})$.

1	1	2	-0.941979	0.506829
1	2	2	-0.767590	0.508578
1	3	2	-0.080773	0.520735
1	4	2	0.461144	0.535310
1	5	2	0.331417	0.530810
1	6	3	1.669464	0.628789
:				
1	2939	2	-0.605034	0.511072
1	2940	2	0.521121	0.537698
1	2941	3	1.493746	0.609831
1	2942	3	1.064398	0.570134

medians but differ in their minima and maxima. This relationship among the minima and maxima is expected given that MAP $\hat{\theta}$ s exhibit greater regression toward the prior's mean (i.e., less variability) than do EAP $\hat{\theta}$ s. (The SDs are 0.818 and 0.841 for the MAP and EAP $\hat{\theta}$ s, respectively.) Our minima and maxima correspond to $X = 0$ and $X = 16$, respectively. This is easily shown using our zero-variance response vectors and mirt's response.pattern argument (e.g., `fscores(. . . , method = "EAP", response.pattern = c(2,2,2,2,2,2,2,2), . . .)`) to obtain the corresponding EAP and MAP $\hat{\theta}$ s. For a perfect response vector ($X = 16$) our EAP $\hat{\theta} = 2.108$ is greater than our MAP $\hat{\theta} = 2.0347$ (i.e., the maxima). Similarly, for a zero response vector ($X = 0$) our EAP $\hat{\theta} = -2.134$ is less than our MAP $\hat{\theta} = -2.052$ (i.e., the minima). Of course and as shown, for zero-variance response vectors MLE will not yield finite $\hat{\theta}$ s (e.g., `fscores(. . . , method = "ML", response.pattern = c(2,2,2,2,2,2,2,2), . . .)`).

Conceptual Development of the Graded Response Model

The PC and GPC models approach the graded scoring of an item by using a series of dichotomous models that govern the probability of responding in one category versus the next adjacent category; the dichotomous models underlying the PC and GPC models are the Rasch and 2PL models, respectively. As an example, and given the category scores for the item "(6/3) + 2 = ?" (i.e., 0, 1, 2 for incorrect, partially correct, and correct responses, respectively), we can focus on the probability of obtaining a category score of 0 versus a score of 1 or the probability of obtaining a score of 1 versus a score of 2. In this fashion we have a series of dichotomous choices. However, this is not the only approach one could use.

An alternative method is used by Samejima (1969) in her extension of Thurstone's method of successive intervals (Masters, 1982). In her approach, there is a boundary above which a person is expected to obtain certain category score(s) as opposed to lower category score(s). The important distinction from the PC/GPC models is in the plurality of the choice. For instance, given the three category scores for the item "(6/3) + 2 = ?,"

TABLE 8.5. mirt Session for the GPC Model Calibration of the Alike Data

```

> # This is a continuation of the session from Table 7.12

> gpcm = mirt(alikedata,1,'gpcm',SE=T,SE.type='Fisher')
   Iteration: 29, Log-Lik: -19646.635, Max-Change: 0.00009

   Calculating information matrix...

> gpcm
   Call:
   mirt(data = alikedata, model = 1, itemtype = "gpcm", SE = T,
         SE.type = "Fisher")

   Full-information item factor analysis with 1 factor(s).
   Converged within 1e-04 tolerance after 29 EM iterations.
   mirt version: 1.30
   M-step optimizer: BFGS
   EM acceleration: Ramsay
   Number of rectangular quadrature: 61
   Latent density type: Gaussian

   Information matrix estimated with method: Fisher
   Condition number of information matrix = 53.59121
   Second-order test: model is a possible local maximum

   Log-likelihood = -19646.64
   Estimated parameters: 24
   AIC = 39341.27; AICc = 39341.68
   BIC = 39484.95; SABIC = 39408.7
   G2 (6536) = 3071.35, p = 1
   RMSEA = 0, CFI = NaN, TLI = NaN

> M2(gpcm,CI=0.95)
      M2 df p      RMSEA  RMSEA_2.5 RMSEA_97.5      SRMSR
      stats 121.2623 12 0 0.05564127 0.04515304 0.06654204 0.04602542
          TLI      CFI
      stats 0.9018448 0.9411069

> anova(gpcm,pcm)    # model 2= gpcm, model 1=pcm
   Model 1: mirt(data = alikedata, model = ConstDiscr, itemtype = "gpcm",
                  SE = T, SE.type = "Fisher")
   Model 2: mirt(data = alikedata, model = 1, itemtype = "gpcm", SE = T,
                  SE.type = "Fisher")

      AIC      AICc      SABIC      HQ      BIC      logLik      X2 df p
      1 39404.53 39404.73 39452.29 39441.17 39506.30 -19685.26      NaN NaN NaN
      2 39341.27 39341.68 39408.70 39393.00 39484.96 -19646.63 77.254    7  0

> coef(gpcm,simplify=TRUE,IRTpars=TRUE)
$items
      a      b1      b2
I1 1.299 -1.542 -1.721
I2 1.059 -0.609 -1.260
I3 0.842 -1.241 -0.280
I4 0.673 -0.323  0.752
I5 0.867  1.815 -1.390
I6 0.625 -0.255  3.862
I7 0.881  1.810  0.035
I8 0.662  1.874  1.781

```

(continued)

TABLE 8.5. (*continued*)

```

$means
F1
0

$cov
F1
F1 1

> mean(coef(gpcm,simplify=T,IRTpars=T)$'items'[,1]) # calc average discr est
[1] 0.8637329

> marginal_rxx(gpcm)
[1] 0.7050771
> plot(gpcm, type = 'trace', theta_lim=c(-4,4)) # produces Figure 8.7

> # obtain person MAP estimates & display first 6 cases
> head((peoplegpcmMAP=fscores(gpcm,method="MAP",full.scores=T,full.scores.SE=T)),6)
      F1      SE_F1
[1,] -0.94161642 0.5068806
[2,] -0.76761489 0.5086084
[3,] -0.08067961 0.5207246
[4,]  0.46127315 0.5352998
[5,]  0.33149986 0.5307964
[6,]  1.66945238 0.6287934

> tail(peoplegpcmMAP,4)
      F1      SE_F1
[2939,] -0.6050650 0.5110873
[2940,]  0.5211951 0.5376866
[2941,]  1.4937927 0.6098380
[2942,]  1.0644211 0.5701296

> fscores(gpcm,method="MAP",full.scores=T,full.scores.SE=T,returnER=T) # emp rxx
      F1
0.7012428

> # obtain person EAP estimates & display first 6 cases
> head((peoplegpcmEAP=fscores(gpcm,method="EAP",full.scores=T,full.scores.SE=T),6)
      F1      SE_F1
[1,] -0.94322274 0.5151800
[2,] -0.76445329 0.5142377
[3,] -0.06876519 0.5238277
[4,]  0.48398344 0.5429810
[5,]  0.35073813 0.5372667
[6,]  1.73525048 0.6373469

> tail(peoplegpcmEAP,4)
      F1      SE_F1
[2939,] -0.5989120 0.5149087
[2940,]  0.5456954 0.5459018
[2941,]  1.5542620 0.6198112
[2942,]  1.1088610 0.5813234

> fscores(gpcm,method="EAP",full.scores=T,full.scores.SE=T,returnER=T) # emp rxx
      F1
0.7068777

> cor(peoplegpcmEAP[,1],peoplegpcmMAP[,1]) # correlation MAP w/ EAP
[1] 0.9999584

```

(continued)

TABLE 8.5. (continued)

```

> sd(peoplegpcmMAP[,1])
[1] 0.8176

> sd(peoplegpcmEAP[,1])
[1] 0.8408505

> summary(peoplegpcmMAP)
      F1           SE_F1
Min. :-2.05235   Min. :0.5066
1st Qu.:-0.54844  1st Qu.:0.5149
Median : 0.03740  Median :0.5245
Mean   :-0.01388  Mean  :0.5330
3rd Qu.: 0.53211  3rd Qu.:0.5430
Max.   : 2.03470  Max.  :0.6706

> summary(peoplegpcmEAP)
      F1           SE_F1
Min. :-2.1338232  Min. :0.5142
1st Qu.:-0.5414926 1st Qu.:0.5191
Median : 0.0508892  Median :0.5306
Mean   : 0.0000608  Mean  :0.5407
3rd Qu.: 0.5569519  3rd Qu.:0.5525
Max.   : 2.1080707  Max.  :0.6752

> # Comparison of MAP & EAP for zero-response vectors;
> #   use of response.pattern argument; regression towards the mean effect
> fscores(gpcm,method="EAP", response.pattern=c(2,2,2,2,2,2,2), full.scores=T,
  full.scores.SE=T) # EAP
      I1 I2 I3 I4 I5 I6 I7 I8      F1       SE_F1
[1,] 2 2 2 2 2 2 2 2 2.108071 0.6752317

> fscores(gpcm,method="MAP", response.pattern=c(2,2,2,2,2,2,2), full.scores=T,
  full.scores.SE=T) # MAP
      I1 I2 I3 I4 I5 I6 I7 I8      F1       SE_F1
[1,] 2 2 2 2 2 2 2 2 2.0347 0.6706245

> fscores(gpcm,method="EAP", response.pattern=c(0,0,0,0,0,0,0,0), full.scores=T,
  full.scores.SE=T) # EAP
      I1 I2 I3 I4 I5 I6 I7 I8      F1       SE_F1
[1,] 0 0 0 0 0 0 0 0 -2.133823 0.5990291

> fscores(gpcm,method="MAP", response.pattern=c(0,0,0,0,0,0,0,0), full.scores=T,
  full.scores.SE=T) # MAP
      I1 I2 I3 I4 I5 I6 I7 I8      F1       SE_F1
[1,] 0 0 0 0 0 0 0 0 -2.052349 0.5740558

> # MLE person location estimates
> fscores(gpcm,method="ML", response.pattern=c(2,2,2,2,2,2,2), full.scores=T,
  full.scores.SE=T) # MLE
      I1 I2 I3 I4 I5 I6 I7 I8      F1       SE_F1
[1,] 2 2 2 2 2 2 2 2 Inf     NA
> fscores(gpcm,method="ML", response.pattern=c(0,0,0,0,0,0,0,0), full.scores=T,
  full.scores.SE=T) # MLE
      I1 I2 I3 I4 I5 I6 I7 I8      F1       SE_F1
[1,] 0 0 0 0 0 0 0 0 -Inf    NA

```

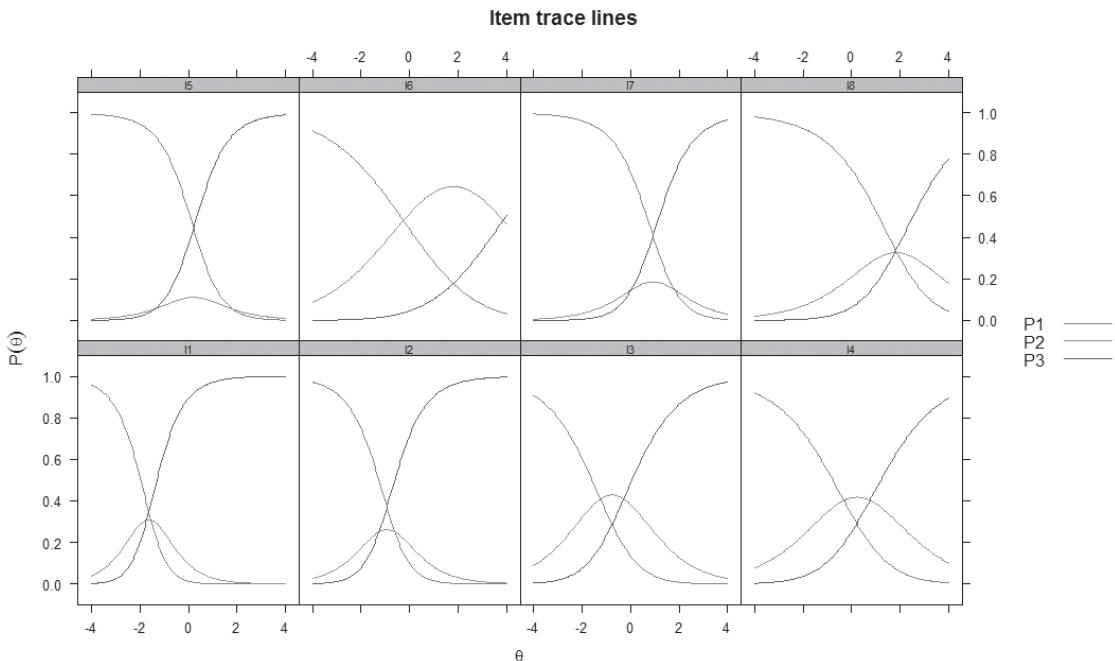


FIGURE 8.7. ORFs for all alike reasoning exam items, GPC model.

we can focus on the probability of obtaining a score of 1 *or higher* versus a score of 0, or the probability of obtaining a score of 2 versus a score of 0 or 1. In effect, the polytomous scores have been turned into a series of *cumulative* comparisons (i.e., below a particular category versus *at and above* this category). We can then trace each of these cumulative probabilities with a dichotomous model. This is the approach used in the *graded response* (GR) model (Samejima, 1969, 2010).

The GR model specifies the probability of a person responding with category score x_j *or higher* versus responding in lower category scores. Stated more generally, the GR model specifies the probability of a person responding in category k or higher versus responding in categories lower than k . (Because the GR model is applicable to situations where partial credit scoring is used as well as Likert response data, we use the terms *category scores* and *categories* interchangeably in the following.) As is the case with the PC model, responses to item j are categorized into $m_j + 1$ categories, where higher categories indicate more of the latent trait. Associated with each of item j 's response categories is a category score, x_j , with integer values $0, 1, \dots, m_j$. According to the GR model, the probability of obtaining x_j *or higher* is given by

$$P_{x_j}^*(\theta) = \frac{e^{\alpha_j(\theta - \delta_{x_j})}}{1 + e^{\alpha_j(\theta - \delta_{x_j})}}, \quad (8.3)$$

where θ is the latent trait, α_j is the discrimination parameter for item j , δ_{x_j} is the *category boundary location* for category score x_j , and $x_j = \{0, 1, \dots, m_j\}$; also see Samejima (1973b).

Alternatively, the category boundary location may be viewed as the boundary between categories k and $k - 1$ (Masters, 1982).² As is the case with the PC and GPC models, the number of categories may vary across items. In the GR model, the δ_{x_j} s are always in increasing order and there are m_j category boundary locations for item j . For simplicity of presentation $P_{x_j}^*$ is used for $P_{x_j}^*(\theta)$ in the following.

The model in Equation 8.3 may be recognized as, in essence, the 2PL model.³ As such, the GR model is the successive application of the 2PL model to an ordered series of bifurcated responses (e.g., 0 vs. 1, 2; 0, 1 vs. 2). Generally speaking, the 2PL model specifies the probability, p_j , of the occurrence of event b , and $(1 - p_j)$ specifies the probability of b 's complementary event \bar{b} (e.g., $b = 1$ and $\bar{b} = 0$); the events \bar{b} and b are mutually exclusive and jointly exhaustive. In the GR model context, the events \bar{b} and b reflect subsets of the x_j (set of) responses. For example, for $m_j = 3$ we have $x_j = \{0, 1, 2, 3\}$. For $x_j = 0$ we have P_0^* with $\bar{b} = \{0, 1, 2, 3\}$ and b is the null set, for $x_j = 1$ we have P_1^* with $\bar{b} = \{0\}$ and $b = \{1, 2, 3\}$, for $x_j = 2$ we have P_2^* with $\bar{b} = \{0, 1\}$ and $b = \{2, 3\}$, and for $x_j = 3$ we have P_3^* with $\bar{b} = \{0, 1, 2\}$ and $b = \{3\}$. Each $P_{x_j}^*$, except for P_0^* , is given by the 2PL model using the x_j category boundary location. By definition, the probability of responding in the lowest category 0 or higher is 1.0 (i.e., $P_0^* \equiv 1$) and the probability of responding in category $m_j + 1$ or higher is 0.0 (e.g., for this item $P_4^* \equiv 0$). In other words, the definition for P_0^* states that the response has to be within one of the categories, whereas the latter definition (e.g., P_4^* or generally, $P_{m_j+1}^*$) states that the probability of responding beyond the highest category is zero.

Given the foregoing, it is not surprising that when $P_{x_j}^*$ is graphed as a function of θ one obtains an ogive. The plot of these cumulative probabilities, sometimes referred to as *category boundary curves*, *cumulative probability curves*, *category characteristic curves* (Dodd, 1984), or *boundary characteristic curves* (Baker, 1992), is similar to a series of IRFs with lower and upper asymptotes of 0.0 and 1.0, respectively. An example of these category boundary curves for an item with three-category scores (i.e., $x_j = \{0, 1, 2\}$; $m_j = 2$) and $\alpha = 1.5$, $\delta_{1j} = -1.0$, and $\delta_{2j} = 1.0$ is shown in Figure 8.8. As can be seen, P_1^* specifies the probability of a score of 0 versus 1 or 2, and P_2^* indicates the probability of a score of 0 or 1 versus 2. Figure 8.8 also shows that the point of inflection of a boundary curve is located at δ_{x_j} . As such, the probability of obtaining a category score x_j or higher is 0.50 at δ_{x_j} . The slopes of the boundary curves at δ_{x_j} are proportional to α_j .

As may be clear from the preceding discussion, the model in Equation 8.3 specifies the probability, $P_{x_j}^*$, of an individual obtaining category score x_j or higher on item j , not the probability of obtaining a specific category score or responding in a particular category, p_{x_j} . To calculate the probability of an individual obtaining a particular category score x_j or responding in a particular category k , p_k , we must take the difference between the cumulative probabilities ($P_{x_j}^*$) for adjacent category sets. That is,

$$p_k = P_k^* - P_{k+1}^* , \quad (8.4)$$

where P_k^* is $P_{x_j}^*$ from Equation 8.3; note the use of lowercase “ p ” to indicate the probability of responding in a particular category and (capital) “ P^* ” to indicate cumulative probabilities.

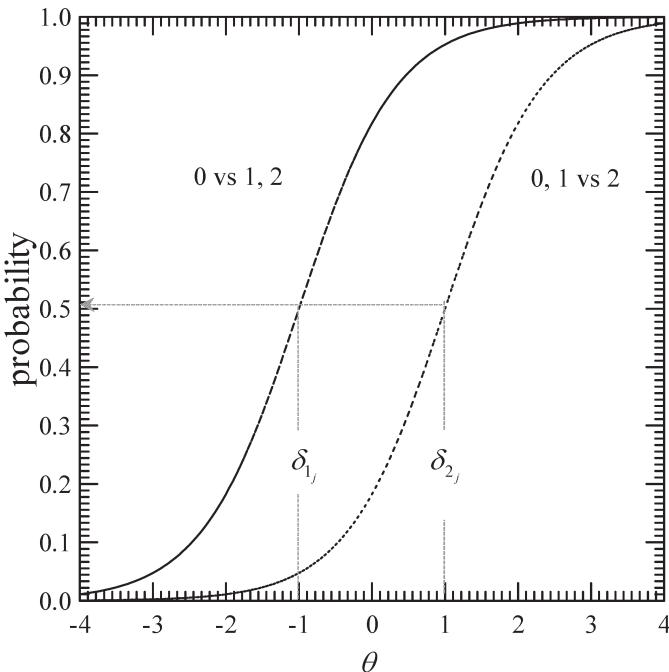


FIGURE 8.8. Category boundary curves for a three-category response item with $\alpha_j = 1.5$, $\delta_{1j} = -1.0$, and $\delta_{2j} = 1.0$.

By way of example, assume that an item has three response categories (i.e., $x_j = \{0, 1, 2\}$; $m_j = 2$). By definition, P_0^* is equal to 1 (i.e., $P_0^* \equiv 1.0$) and specifies the probability of responding in category 0, 1, or 2. The term P_1^* is the probability of responding in category 1 or 2 rather than in category 0 (i.e., $k = 1$ or $x_j = 1$). Also, P_2^* is the probability of responding in category 2 rather than in category 0 or 1 (i.e., $k = 2$ or $x_j = 2$), and $P_3^* \equiv 0$. Therefore, the probability of responding in category 0 (i.e., $x_j = 0$) is given by

$$p_0 = P_0^* - P_1^* = p(x_j = \{0, 1, 2\} | \theta) - p(x_j = \{1, 2\} | \theta) = 1.0 - \frac{e^{\alpha_j(\theta - \delta_{1j})}}{1 + e^{\alpha_j(\theta - \delta_{1j})}}$$

The probability of responding in category 1 (i.e., $x_j = 1$) is

$$p_1 = P_1^* - P_2^* = p(x_j = \{1, 2\} | \theta) - p(x_j = 2 | \theta) = \frac{e^{\alpha_j(\theta - \delta_{1j})}}{1 + e^{\alpha_j(\theta - \delta_{1j})}} - \frac{e^{\alpha_j(\theta - \delta_{2j})}}{1 + e^{\alpha_j(\theta - \delta_{2j})}}$$

and the probability of responding in category 2 (i.e., $x_j = 2$) equals

$$p_2 = P_2^* - P_3^* = p(x_j = 2 | \theta) - p(x_j > 2 | \theta) = \frac{e^{\alpha_j(\theta - \delta_{2j})}}{1 + e^{\alpha_j(\theta - \delta_{2j})}} - 0.$$

The sum of the p_k s across the response options for a fixed value of θ is 1.0 (i.e., $\sum_{x_j=0}^m p_k = 1$).⁴

The plot of the p_{x_j} s as a function of θ (i.e., the ORFs; operating characteristic functions) for the item in Figure 8.8 are shown in Figure 8.9. Note the δ_{x_j} s do not necessarily correspond to the points of intersection of adjacent ORFs. Rather, for the lowest and highest categories, the δ_{x_0} (p_0) and δ_{x_m} (p_2), respectively, correspond to where the probability of responding in the given category is 0.5. As a second example, Figure 8.10 contains the ORFs for a four-response category item with $\alpha = 1.5$, $\delta_{1j} = -1.0$, $\delta_{2j} = 1.4$, and $\delta_{3j} = 2$. The probability of responding $x_j = 0$ (p_0) is 0.5 at δ_{1j} and the probability of responding $x_j = 3$ (p_3) is 0.5 at δ_{3j} .

Because with the GR model we can only have sequentially ordered δ_{x_j} s, it is not possible to inspect the δ_{x_j} s for an indication of which categories are unlikely to be chosen. Thus, for the GR model it is necessary to plot the ORFs to determine which categories (if any) are less likely to be chosen. For example, Figure 8.11 contains the ORFs for an item with the same slope as the item in Figure 8.9, but with different category boundary locations. As can be seen, although the δ_{x_j} s are in sequence, the corresponding ORFs show that the item is behaving primarily as a dichotomous item, with a category score of 1 (p_1) being less likely than a category score of 0 (p_0) or 2 (p_2).

It can be seen from Figure 8.8 that the boundary curves are parallel. This parallelism reflects the assumption in Equation 8.3 that discrimination is constant across the graded response categories (i.e., α_j is constant within an item across boundaries, but not necessarily across items). This is sometimes referred to as the *homogeneous* GR model

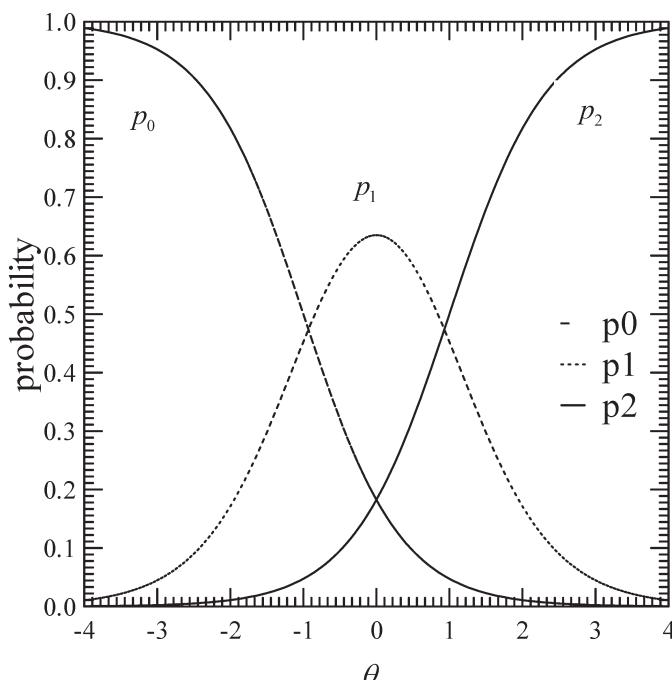


FIGURE 8.9. ORFs for a three-category response item with $\alpha_j = 1.5$, $\delta_{1j} = -1.0$, and $\delta_{2j} = 1.0$.

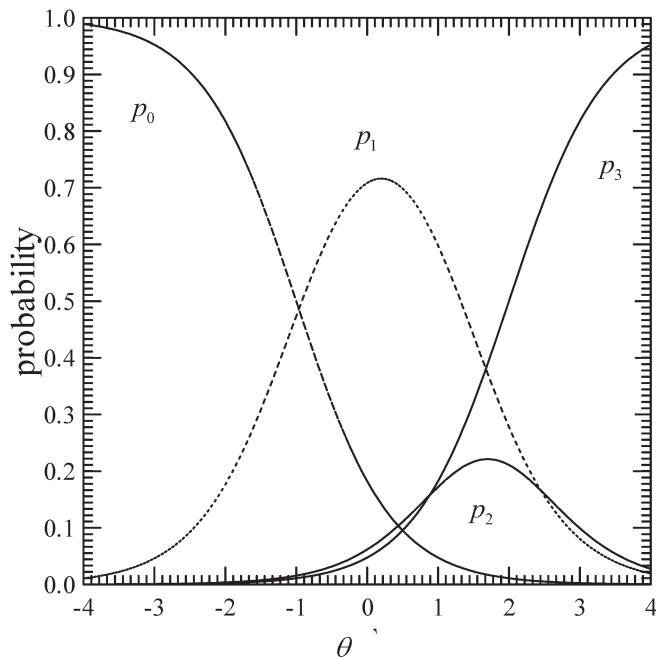


FIGURE 8.10. ORFs for a four-response category item with $\alpha_j = 1.5$, $\delta_{1j} = -1.0$, $\delta_{2j} = 1.4$, and $\delta_{3j} = 2$.

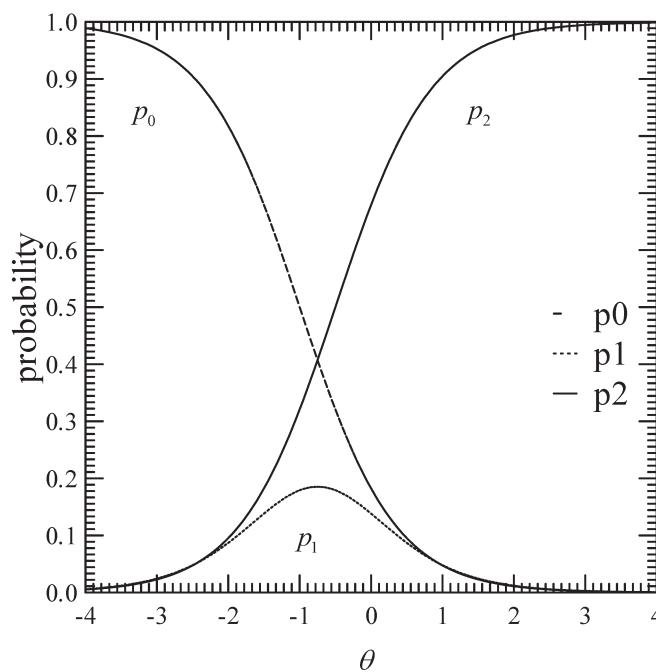


FIGURE 8.11. ORFs for a three-category response item with $\alpha_j = 1.5$, $\delta_{1j} = -1.0$, and $\delta_{2j} = -0.5$.

and is the model most typically used. In contrast, α_j may be allowed to vary across the graded response categories as well as across items. This second variant may be called the *heterogeneous GR model*; see Samejima (1969, pp. 19–20). These models may be seen as special cases of the *continuous response model* (Samejima, 1973b).

As is true with the PC, RS, and GPC models, the GR model can be used for both cognitive and attitude measurement. For instance, Samejima (1969) applied the GR model to the analysis of nonverbal reasoning ability, Koch (1983) used the model for the analysis of teacher attitudes to the communication skills of school administrators, and Steinberg (2001) examined the context effects of anger experience and anger expression questions through the GR model. Dodd (1984) provided a comparative analysis of the GR and PC models for attitude measurement. She found that, in general, using the PC and GR models for attitude assessment produced results that were highly related to traditional analysis methods while also providing the advantages of (1) knowing the precision of the measurement attained by the attitude scale for each individual and (2) facilitating the development of sample-independent scales. The GR model has also been applied to indirect measures of writing assessment (Ackerman, 1986) and to computerized adaptive testing (CAT; Samejima, 1976).

How Large a Calibration Sample?

In a study of parameter recovery in the GR model, Reise and Yu (1990) found that sample size did not affect estimation of the person location parameter but did affect estimation of the item parameters. They recommended that at least 500 respondents were needed to achieve an adequate calibration with the GR model. Their study was conducted with 25 five-response category items, and, therefore, their guidelines (strictly speaking) are appropriate only for instruments of this length.

As is the case with the PC and RS models, because of the interaction of the distribution of the respondents across the response categories, as well as across the items, it is difficult to arrive at a hard-and-fast guideline that would be applicable in all situations. However, for guidance we provide very rough guidelines. Assuming MMLE, a symmetric θ distribution, and that the respondents distribute themselves across the response categories in reasonable numbers, we suggest that the minimum sample size be, say 500. This value is a quasi-lower bound that serves to address the issues of fit analysis, dimensionality assessment, ensuring that there are respondents in each category, difficulty in estimating discrimination parameters, and so on. As previously stated, it may be anticipated that there is a sample size, say 1,200, at which one reaches, practically speaking, a point of diminishing returns in terms of improvement in estimation accuracy. (As previously stated, this 1,200 should not be interpreted as an upper bound.) If one adopts a sample size ratio for sample size determination (e.g., five persons for every parameter estimated), then it is probably more useful closer to the lower bound than to the 1,200 value (i.e., when the sample size is large, then the sample size ratio becomes less important). As previously mentioned, these sample size guidelines should not be interpreted as hard-and-fast rules, and these suggestions are tempered by the purpose

of the administration (e.g., survey, establishing norms, equating, item pool development/maintenance), the estimation approach, the application's characteristics (e.g., distribution and range of transition/category boundary locations, instrument length, latent distribution shape), ancillary technique sample size requirements, the use of a prior distribution for estimating α , and the amount of missing data.

Information for Graded Data

As discussed in Chapter 7, with polytomous models, it is possible to determine the amount of information provided by each response category. When one scores an item in a graded way, the information function for the graded response is (Samejima, 1969)

$$I_{x_j}(\theta) = \left\{ -\frac{\partial^2 \ln p_k}{\partial \theta^2} \right\} p_k \quad (8.5)$$

That is, each graded response (potentially) contributes some information for estimating a person's location. The sum of these option information functions, $I_{x_j}(\theta)$, equals the item's information, $I_j(\theta)$. An item's information is

$$I_j(\theta) = \sum_{x_j=0}^{m_j} I_{x_j}(\theta) = \sum_{x_j=0}^{m_j} \frac{(p'_{x_j})^2}{p_{x_j}} \quad (8.6)$$

When applied to dichotomous data, Equation 8.6 simplifies to the item information formula presented in Equation 2.16.

As seen in previous chapters, the sum of the item information yields the instrument's total information

$$I(\theta) = \sum_{j=1}^L I_j(\theta). \quad (8.7)$$

In the case of the GR model and given Equation 8.4, the option information function is

$$I_{x_j}(\theta) = \frac{(p'_{x_j})^2}{p_{x_j}} = \frac{\left[P'_{x_j}^* - P'_{x_{j+1}}^* \right]^2}{P_{x_j}^* - P_{x_{j+1}}^*}, \quad (8.8)$$

where $P_{x_j}^*$ is given by Equation 8.3 and $P'_{x_j}^*$ is its first derivative. As mentioned above, $P_{x_j}^*$ is, in effect, the 2PL model and its first derivative of the 2PL model is $\alpha_j p_j (1 - p_j)$. Consequently, we have that $P_{x_j}^* P'_{x_j}^* = \alpha_j P_{x_j}^* (1 - P_{x_j}^*)$.

By way of an example, let $m_j = 2$, and for ease of presentation, let $\varphi_{x_j} = [1 + \exp(\alpha_j(\theta - \delta_{x_j}))]^2$. In the following, we first concentrate on the bracketed term in the numerator of Equation 8.8 and then proceed to calculate the option and item information.

When $x_j = 0$, we have that $P_0^* \equiv 1.0$ and

$$\varphi_1 = [1 + \exp(\alpha_j(\theta - \delta_{x_j}))]^2$$

so that

$$p'_0 = P_0^{**} - P_1^{**} = 0 - \alpha_j \left[\frac{\exp(\alpha_j(\theta - \delta_1))}{\varphi_1} \right] = -\alpha_j \frac{\exp(\alpha_j(\theta - \delta_1))}{\varphi_1}.$$

For the category score $x_j = 1$ we have

$$\varphi_2 = (1 + \exp(\alpha_j(\theta - \delta_2)))^2$$

and

$$\begin{aligned} p'_1 &= P_1^{**} - P_2^{**} = \alpha_j \left[\frac{\exp(\alpha_j(\theta - \delta_1))}{\varphi_1} \right] - \alpha_j \left[\frac{\exp(\alpha_j(\theta - \delta_2))}{\varphi_2} \right] \\ &= \alpha_j \left[\left(\frac{\exp(\alpha_j(\theta - \delta_1))}{\varphi_1} \right) - \left(\frac{\exp(\alpha_j(\theta - \delta_2))}{\varphi_2} \right) \right]. \end{aligned}$$

For the last category score, $x_j = 2$, we have that $P_3^{**} \equiv 0$. Therefore,

$$p'_2 = P_2^{**} - P_3^{**} = \alpha_j \left[\frac{\exp(\alpha_j(\theta - \delta_2))}{\varphi_2} \right] - 0 = \alpha_j \left[\frac{\exp(\alpha_j(\theta - \delta_2))}{\varphi_2} \right].$$

To obtain the item's information we square each of these first derivatives, divide it by the probability of responding in the corresponding category, and then sum the quotients

$$I_j(\theta) = \sum_{x_j=0}^{m_j} \frac{(p'_{x_j})^2}{p_{x_j}} = \frac{(p'_0)^2}{p_0} + \frac{(p'_1)^2}{p_1} + \frac{(p'_2)^2}{p_2},$$

where each quotient is the option information function for the corresponding x_j .

Figure 8.12 contains the option and item information functions for the item shown in Figures 8.8 and 8.9. Given Equation 8.6, it is not surprising that the item information for this item is greater than the individual option information functions. These individual option information functions vary not only in their individual contributions to the item information, but also in how their information is distributed across θ . For example, the option information functions for $x_j = 0$ and 2 are unimodal, whereas for $x_j = 1$ the function is bimodal with a minimum around 0.0. Because the maxima of the option information functions for $x_j = 0$ and 2 are in the vicinity of 0.0, this item is able to provide a somewhat uniform amount of information for estimating individuals between roughly -1.5 and 1.5. The distribution of information is a function of the distances between δ_{x_j} s (Samejima, 1969).

Samejima (1969) shows that there is an increase in item information if a response category is added between two adjacent categories. Therefore, the amount of item infor-

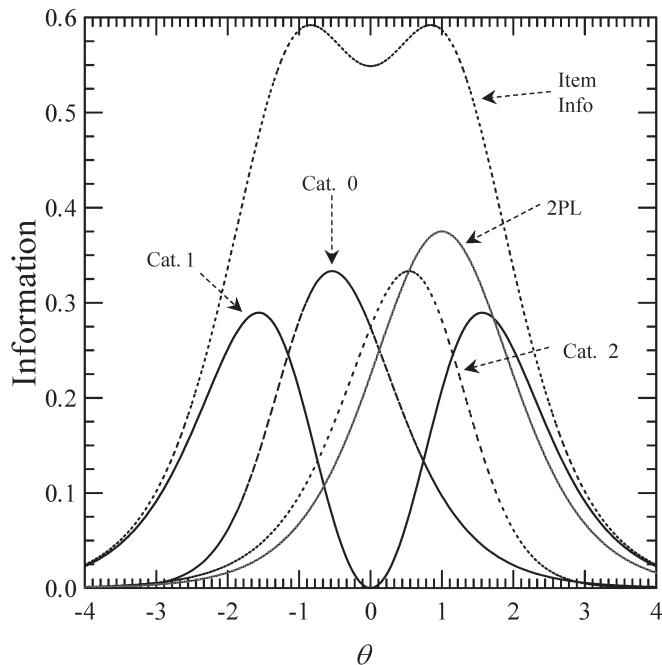


FIGURE 8.12. Category and item information functions for a three-category response item with $\alpha_j = 1.5$, $\delta_{1j} = -1.0$, and $\delta_{2j} = 1.0$ and a dichotomous item ($\alpha_j = 1.5$ and $\delta_{2j} = 1$).

mation available by treating an item in a polytomous graded fashion is at least equal to, and more likely greater than, the amount of item information available when the item is scored in a dichotomous fashion. This is demonstrated in Figure 8.12. The 2PL line shows the information available when this item is dichotomously scored (i.e., $x_j = 0$ or 1 is considered to be incorrect). As can be seen, the graded scoring results in greater item information than when the item is dichotomously scored. Moreover, this information is over a broader range of θ than when the item is scored dichotomously. In general, polytomously scored items tend to provide greater information toward the lower end of the continuum than when dichotomously scored. For our example item, we see that the graded scoring results in greater information, particularly below the item's location of 1, than when the item is dichotomously scored.

Metric Transformation, GPC and GR Models

To transform the item discrimination (or its estimate) one uses $\alpha^* = \alpha/\zeta$ and for the category location parameters (or their estimates) one uses $\delta_{x_j}^* = \zeta(\delta_{x_j}) + \kappa$. The θ metric can be transformed to another metric (e.g., an “attitude toward condoms” metric that has some arbitrary mean and unit) by $\theta^* = \zeta(\theta) + \kappa$.

Additionally, we can convert the person locations (or their estimates) to their cor-

responding expected trait scores through the sum of the expected item responses. For the GPC model where 1 indicates the lowest response category we use

$$T = \sum_{k=1}^L \left(\sum_{j=1}^{m_j} kp(x_{jk} | \theta) \right). \quad (8.9)$$

T has a range from L to $\sum_{j=1}^{m_j}$ and the bracketed term is the expected item response.

For the GR model with 0 representing the lowest response category we use

$$T = \sum_{k=0}^L \left(\sum_{j=1}^{m_j} kp(x_{jk} | \theta) \right) \quad (8.10)$$

and the bracketed term is the expected item response. In this case, T's range is from 0 to $\sum_{j=1}^{m_j}$. When 1 represents the lowest response category (e.g., as might be used with a Likert response scale), the limits of summation for the inside-the-brackets summation would be 1 and $(m_j + 1)$; the range of T is L to $\sum_{j=1}^{m_j} + L$. The expected item response from either Equation 8.9 or 8.10 can be used with Q_3 to examine conditional dependence.

Example: Application of the GR Model to an Attitudes Toward Condoms Scale, MMLE, flexMIRT

Several programs are available for GR model estimation (e.g., flexMIRT, mirt, SAS proc irt). We calibrate our data with flexMIRT followed by mirt. In Chapter 7 we apply the RS model to survey data from an attitudes toward condoms instrument. We now revisit these data. Recall that our instrument contains six statements people have made about condoms. The respondents are asked how much they agree with each of the statements on a 4-point Likert response scale (1 = "strongly disagree," 2 = "disagree more than I agree," 3 = "agree more than I disagree," 4 = "strongly agree"). Because the six statements are "negatively" worded, a respondent who strongly agrees with a statement is indicating a less favorable attitude toward condom use. As we did in Chapter 7, we assume that we have evidence supporting the tenability of the unidimensionality assumption.

The flexMIRT command file for this calibration is presented in Table 8.6. Our data contain case identification in the first column followed by the item responses. This is reflected in our Varnames command (Varnames = id,i1-i6;). We specify the case identification variable name (id) with the CaseID command and subsequently identify the variables to be used in the calibration (i.e., Select = i1-i6). For our case calibration we need to re-code our 1-based responses (i.e., 1 = "strongly disagree," 2 = "disagree more than I agree," . . . , 4 = "strongly agree") to be 0-based (i.e., 0 = "strongly disagree," 1 = "disagree more than I agree," . . . , 3 = "strongly agree"). Using the Code command and specifying all six items accomplishes this re-coding. Because the GR model is a primary model, we use keyword GRADED on the Model line to indicate a GR model

TABLE 8.6. Command File for the flexMIRT GR Model Calibration Example

```

<Project>
  Title = "GR calibration, Condoms data, 6 four category items ";
  Description = "Obtain & save item & person parameter estimates";

<Options>
  Mode = Calibration;
  GOF = Extended;
  NumDec = 3;
  savePRM= Yes;
  SCORE= EAP;
  saveSCO= Yes;
  FisherInf= 81, 4.0;
  saveINF= Yes;
  SE= Fisher;

<Groups>
  %OnlyGroup%
  File = "condomsSpcDlmtCase.DAT";           // space delimited file
  Varnames = id,i1-i6;                         // identify case id variable
  CaseID =id;                                  // select variables to calibrate
  Select=i1-i6;
  N = 3473;
  Code(i1-i6) = (1,2,3,4), (0,1,2,3);        // recode responses '1'...'4' to be 0-based
  Ncats(i1-i6) = 4;                            // 4 valid responses
  Model(i1-i6) = Graded(4);                  // specify the GR model, a primary model

<Constraints>

```

calibration. As above, we specify the number of categories per item by using the Ncats keyword.

The output is presented in Table 8.7. As can be seen, convergence is achieved. Our item parameter estimates are found in the Graded Items for . . . table. For example, for the first item we have $\hat{\alpha}_1 = 1.016$, $\hat{\delta}_{11} = -0.555$, $\hat{\delta}_{21} = 0.129$, and $\hat{\delta}_{31} = 1.003$, and for the last item $\hat{\alpha}_6 = 0.768$, $\hat{\delta}_{16} = -1.883$, $\hat{\delta}_{26} = -0.933$, and $\hat{\delta}_{36} = 0.403$. As would be expected, the $\hat{\delta}_{xj}$ s are in sequential order.

If we ignore the philosophical differences underlying the RS and GR models and compare our fit to that of the RS model (Chapter 7, Endnote 19) we find both of our information criteria indicating the GR model reflects a comparatively better fit to the data than the RS model. That is, when we relax the $\alpha = 1$ and constant threshold across items constraints used with the RS model, we obtain better model data fit. Of course, this does not mean that the data do not conform to the RS model sufficiently well that the model is not useful. (Because the RS and GR models are not hierarchically related, we cannot use ΔG^2 to determine if one fits significantly better.) Because the RS model is implemented as a constrained version of the nominal model, we have eight parameters to estimate. In contrast, the GR model is estimating three times as many (see NUMBER OF FREE PARAMETERS: 24 line). These 24 item parameters reflect three- $\hat{\delta}_{xj}$ plus one α_j for each of the six items. With these additional parameters it is not surprising that the GR model fits better than the RS model.

TABLE 8.7. Abridged Output from the flexMIRT GR Model Calibration of the Condoms Data

```

:
GR calibration, Condoms data
Obtain & save item parameter estimates
:
Convergence and Numerical Stability
flexMIRT(R) engine status: Normal termination
First-order test: Convergence criteria satisfied
Condition number of information matrix: 148.1535
Second-order test: Solution is a possible local maximum
:
Number of free parameters: 24
:
Graded Items for Group 1: OnlyGroup
Item Label P#     a      s.e.    b 1    s.e.    b 2    s.e.    b 3    s.e.
  1   i1    4    1.016  0.053 -0.555  0.047  0.129  0.041  1.003  0.059
  2   i2    8    2.092  0.118  0.387  0.029  0.738  0.034  1.131  0.043
  3   i3   12    2.148  0.124  0.504  0.030  0.869  0.036  1.233  0.045
  4   i4   16    0.671  0.044 -1.574  0.109 -0.339  0.059  1.069  0.085
  5   i5   20    0.904  0.052 -0.740  0.057 -0.150  0.044  0.508  0.050
  6   i6   24    0.768  0.047 -1.883  0.113 -0.933  0.070  0.403  0.055
:
Marginal reliability for response pattern scores: 0.67

Statistics based on the loglikelihood of the fitted model:
          -2loglikelihood: 48310.40
          Akaike Information Criterion (AIC): 48358.40
          Bayesian Information Criterion (BIC): 48506.07

Full-information fit statistics of the fitted model:
          Degrees
          G2 of freedom Probability      F0hat      RMSEA
        4771.21           1184      0.0001     1.3738      0.03
          Degrees
          X2 of freedom Probability      F0hat      RMSEA
        10099.13           4071      0.0001     2.9079      0.02
:
The following are the EAP  $\hat{\theta}$ s found in the saved '-sco.txt' file. The format is Group, record's ordinal position in the data file, our case ID, the  $\hat{\theta}$ , and  $s_e(\hat{\theta})$ .

```

1	1 1	-1.536215	0.734371
1	2 2	-1.536215	0.734371
1	3 3	-0.987507	0.684344
1	4 4	-1.536215	0.734371
1	5 5	-1.536215	0.734371
1	6 6	-1.021685	0.670939
:			
1	3470 3470	1.923816	0.606203
1	3471 3471	1.923816	0.606203
1	3472 3472	1.923816	0.606203
1	3473 3473	1.923816	0.606203

Example: Application of the GR Model to an Attitudes Toward Condoms Scale, MMLE, mirt

Table 8.8 shows our mirt session. Because our data file is a csv formatted file, we can use the `read.csv` function to read the data. However, we use the `read.table` function with the `separator` argument (`sep = ","`) in lieu of `read.csv` to demonstrate its use. Moreover, because our data file contains format information on the first line (Format: `id case, i1 - i6`), we use the `skip` argument (`skip = 1`) to ignore the file's first line. Because mirt will produce the same results using 0-based or 1-based responses, there is no need to re-code our responses. After removing the `id` variable, we use the `describe` function from the `Hmisc` package (`describe(condomsdata)`). Our results show the correct number of response categories for each item, that N is correct, and that we do not have any categories with very small or zero frequencies.

To perform our calibration, we specify `graded` as our `itemtype` in our call to the `mirt` function (`grm = mirt(condomsdata, model = 1, itemtype = "graded", SE = T)`) and print the output object, `grm`. Our calibration required 23 iterations to obtain convergence.

Above (e.g., Chapters 3 and 4) we mentioned inspection of the iteration history as part of examination of the results. This examination could be useful in, for example, determining the necessity of increasing the maximum number of iterations to allow convergence. For instance, if the history showed the log likelihood “bouncing” back and forth, then increasing the maximum number of iterations would most likely not be useful. For pedagogical reasons we show obtaining the iteration history. We use the `extract.mirt` function with the `LLhistory` argument to obtain the calibration’s iteration history. The history shows the desired progression of smaller and smaller changes in the log likelihood culminating in a `lnL` of `-24,155.20366`; the 24th iteration has a `lnL` that is less than `0.0001` different (i.e., Converged within `1e-04` tolerance) than that of the 23rd iteration.

Our item parameter estimates are for item 1 $\hat{\alpha}_1 = 1.017$, $\hat{\delta}_{11} = -0.555$, $\hat{\delta}_{21} = 0.128$, and $\hat{\delta}_{31} = 1.002$, for item 2 $\hat{\alpha}_2 = 2.089$, $\hat{\delta}_{12} = 0.387$, $\hat{\delta}_{22} = 0.738$, and $\hat{\delta}_{32} = 1.131$, and so on. Most items show reasonable discrimination capacity. We use the `plot` function to obtain our items ORFs (Figure 8.13). In our call we use the `which.items` argument to order the item graphs to be in standard left-to-right/top-to-bottom layout. In addition, we use the `par.settings` argument to set the use of different line types (`lty = 1:4`). As can be seen, respondents have a tendency to either “strongly agree” or “strongly disagree” to each item. The corresponding item information functions are shown in Figure 8.14 (top panel). It is apparent that items 2 and 3 ($\hat{\alpha}_2 = 2.089$, $\hat{\alpha}_3 = 2.143$) provide substantially more information than the remaining items and item 4 provides the least information for person estimation ($\hat{\alpha}_4 = 0.671$). Overall, the instrument’s total information function (Figure 8.14, bottom panel) shows the scale performs well in estimating individuals in the approximate θ range of 0 to 2.

As above, we obtain our EAP $\hat{\theta}$ s using `fscores` and the corresponding fit information using `personfit`. Given our response scale and the negative wording of the items, our first person is estimated to have a positive attitude ($\hat{\theta}_1 = -1.5367$) toward condom

TABLE 8.8. mirt Session for the GR Model Calibration of the Condoms Data

```

> load mirt & Hmisc
> condomsdata = read.table(file.choose(), sep=",", skip=1, col.names=c("id", paste0("I", 1:6)))
> condomsdata=within(condomsdata, rm(id))

> Hmisc::describe(condomsdata)
condomsdata

 6 Variables      3473 Observations
-----
I1
    n missing distinct      Info      Mean      Gmd
 3473       0         4     0.908     2.388     1.392

Value      1      2      3      4
Frequency  1338   492   602 1041
Proportion 0.385 0.142 0.173 0.300
-----
I2
    n missing distinct      Info      Mean      Gmd
 3473       0         4     0.752     1.847     1.18

Value      1      2      3      4
Frequency  2157   345   315   656
Proportion 0.621 0.099 0.091 0.189
-----
I3
    n missing distinct      Info      Mean      Gmd
 3473       0         4     0.711     1.754     1.098

Value      1      2      3      4
Frequency  2280   344   272   577
Proportion 0.656 0.099 0.078 0.166
-----
I4
    n missing distinct      Info      Mean      Gmd
 3473       0         4     0.924     2.621     1.345

Value      1      2      3      4
Frequency  956    596   728 1193
Proportion 0.275 0.172 0.210 0.344
-----
I5
    n missing distinct      Info      Mean      Gmd
 3473       0         4     0.884     2.581     1.44

Value      1      2      3      4
Frequency  1243   380   440 1410
Proportion 0.358 0.109 0.127 0.406
-----
I6
    n missing distinct      Info      Mean      Gmd
 3473       0         4     0.896     2.875     1.279

Value      1      2      3      4
Frequency  744    451   774 1504
Proportion 0.214 0.130 0.223 0.433
>
> print((grm=mirt(condomsdata, model=1, itemtype="graded", SE=T)))
  Iteration: 23, Log-Lik: -24155.204, Max-Change: 0.00009

  Calculating information matrix...

  Call:
  mirt(data = condomsdata, model = 1, itemtype = "graded", SE = T)

```

(continued)

TABLE 8.8. (continued)

```

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 23 EM iterations.
mirt version: 1.30
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Information matrix estimated with method: Oakes
Condition number of information matrix = 141.0064
Second-order test: model is a possible local maximum

Log-likelihood = -24155.2
Estimated parameters: 24
AIC = 48358.41; AICc = 48358.76
BIC = 48506.07; SABIC = 48429.81
G2 (4071) = 4771.22, p = 0
RMSEA = 0.007, CFI = NaN, TLI = NaN

> # display iteration history
> print((as.data.frame((extract.mirt(grm,'LLhistory')))),digits=10)
  (extract.mirt(grm, "LLhistory"))
  1          -24346.09238
  2          -24257.32861
  3          -24208.03581
  4          -24184.10680
  5          -24169.27831
  6          -24162.60275
  7          -24159.44164
  8          -24157.93437
  9          -24156.76108
 10         -24155.42307
 11         -24155.32346
 12         -24155.26628
 13         -24155.21534
 14         -24155.20997
 15         -24155.20715
 16         -24155.20710
 17         -24155.20686
 18         -24155.20570
 19         -24155.20433
 20         -24155.20419
 21         -24155.20394
 22         -24155.20398
 23         -24155.20366

> coef(grm,simplify=T,IRTpars=T)
  $items
    a      b1      b2      b3
  I1 1.017 -0.555  0.128  1.002
  I2 2.089  0.387  0.738  1.131
  I3 2.143  0.504  0.870  1.233
  I4 0.671 -1.574 -0.339  1.069
  I5 0.904 -0.740 -0.150  0.507
  I6 0.768 -1.883 -0.933  0.402

  $means
  F1
  0

  $cov
    F1
  F1  1

```

(continued)

TABLE 8.8. (continued)

```

> plot(grm, type = 'trace', which.items = c(4,5,6,1,2,3),theta_lim=c(-4,4),
  auto.key=list(points=FALSE,lines=TRUE, columns=4),par.settings = simpleTheme(lty=1:4,1
  wd=2))                                         # all item ORFs (Figure 8.13)

> plot(grm,type="infotrace",theta_lim=c(-4,4))      # all item information (Figure 8.14 top)
> plot(grm,type="info",theta_lim=c(-4,4))           # total information (Figure 8.14 bottom)

> plot(grm,type="score",theta_lim=c(-4,4))          # total characteristic function (Figure 8.15)

> # obtain person location estimates
> peopleGRM=fscores(grm,method="EAP",full.scores=T,full.scores.SE=T)

> mean(peopleGRM[,1])
[1] -0.0003037159

> sd(peopleGRM[,1])
[1] 0.8208128

> peopleGRMFit=personfit(grm,method="EAP")          # obtain person fit info

> peopleGRMnFit=cbind(peopleGRM,peopleGRMFit)        # combine person est & fit info

> head(peopleGRMnFit,6)
      F1      SE_F1      Zh
1 -1.5367232 0.7342694 1.3713116
2 -1.5367232 0.7342694 1.3713116
3 -0.9879395 0.6843462 0.7635174
4 -1.5367232 0.7342694 1.3713116
5 -1.5367232 0.7342694 1.3713116
6 -1.0220257 0.6709094 0.3930765

> tail(peopleGRMnFit,4)
      F1      SE_F1      Zh
3470 1.923777 0.606422 1.32875
3471 1.923777 0.606422 1.32875
3472 1.923777 0.606422 1.32875
3473 1.923777 0.606422 1.32875

> write.csv(peopleGRM, file = "peopleGRM_EAPfit.csv")  # save theta estimates & fit info

```

usage, whereas respondent #3473 is a negative attitude ($\hat{\theta}_{3473} = 1.9238$). Of course, if one wishes to have θ s sign correspond to the attitude interpretation, we could reflect θ 's sign or reverse code the responses prior to calibration. If we wish to report respondent attitudes on the observed (total score) scale, we can use the expected trait scores function to transform $\hat{\theta}$ to T (Figure 8.15). (With our four one-based response categories, our observed scores have a range from 6 to 24 (inclusive).) For example, a respondent with a negative attitude toward condom use, say $\hat{\theta} = 2$, would be expected to have an expected trait score of approximately 22; the exact value can be obtained using Equation 8.9.

Conceptual Development of the Continuous Response Model

When we measure the length of a table using a tape measure, we divide its length into multiples of a unit of measurement. If our tape measure's unit of measurement is 1/16th

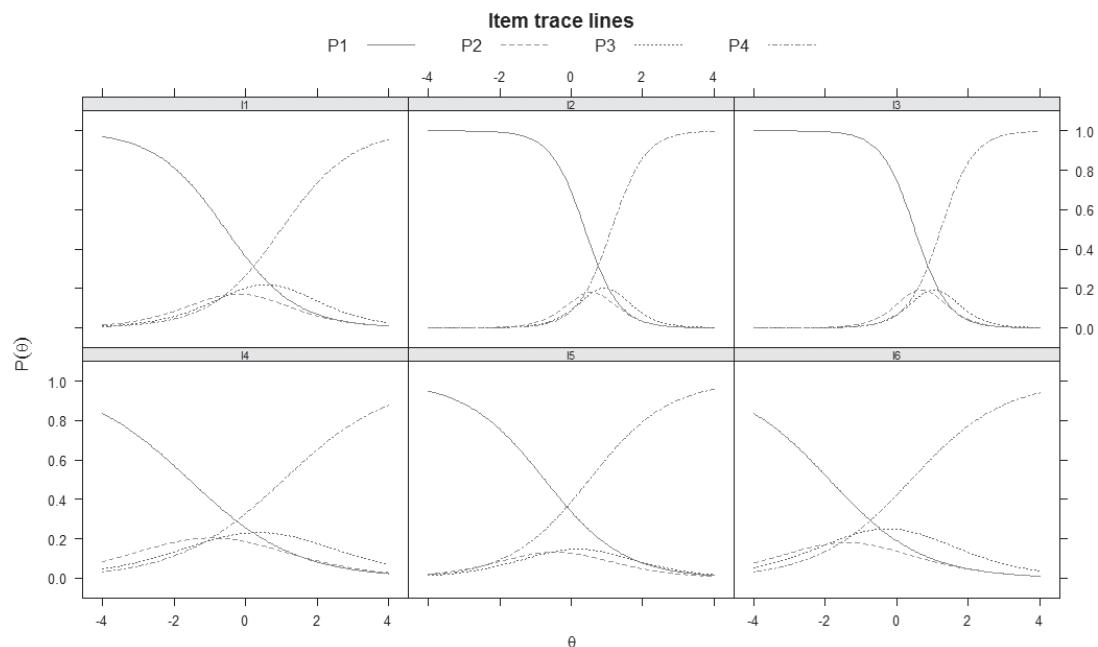


FIGURE 8.13. ORFs for all items.

of an inch (or it's a mm), then the length of our table will be a multiple of a 1/16th of an inch (or of a mm). Although the accuracy of our measurement is determined by our unit, we can use finer gradations of the unit (e.g., 1/32nd, 1/64th) to obtain better approximations of the table's length. Of course, because our table's length is infinitely divisible, we consider length to be continuous. We apply this idea to attitude assessment.

Recall with our Attitudes Toward Condoms scale we used a 4-point Likert response scale ranging from “strongly disagree” to “strongly agree” where our unit is a point and our responses are multiples of that point. (We label our response scale in terms of the maximum multiple.) As such, our category score, x_j , can take on values of 1, 2, 3, and 4. The number of category scores is $m_j = 4$. (Note the italicized “ m “ indicates the number of category scores, not the number of transition locations as seen with, say, the PC model.) Of course, we could have used a 5- or a 7-point scale. In fact, we could increase the number of discrete response categories to be, say a 100 (i.e., $m_j = 100$), so that each response category has a value that is a multiple of 1/100th. Let us rescale our response scale to be [0, 1]. That is, our endpoints are 0 = “strongly disagree” and 1 = “strongly agree.” Therefore, the category immediately next to “strongly disagree” would have a value of 1/100th, the following one would be 2/100th, and so on; 1/100th is our unit. An item response format that can be used in this context is the graphic response scale (Hayes & Patterson, 1921; Freyd, 1923) without descriptors along the line.⁵ Three example items using a *continuous rating scale* are

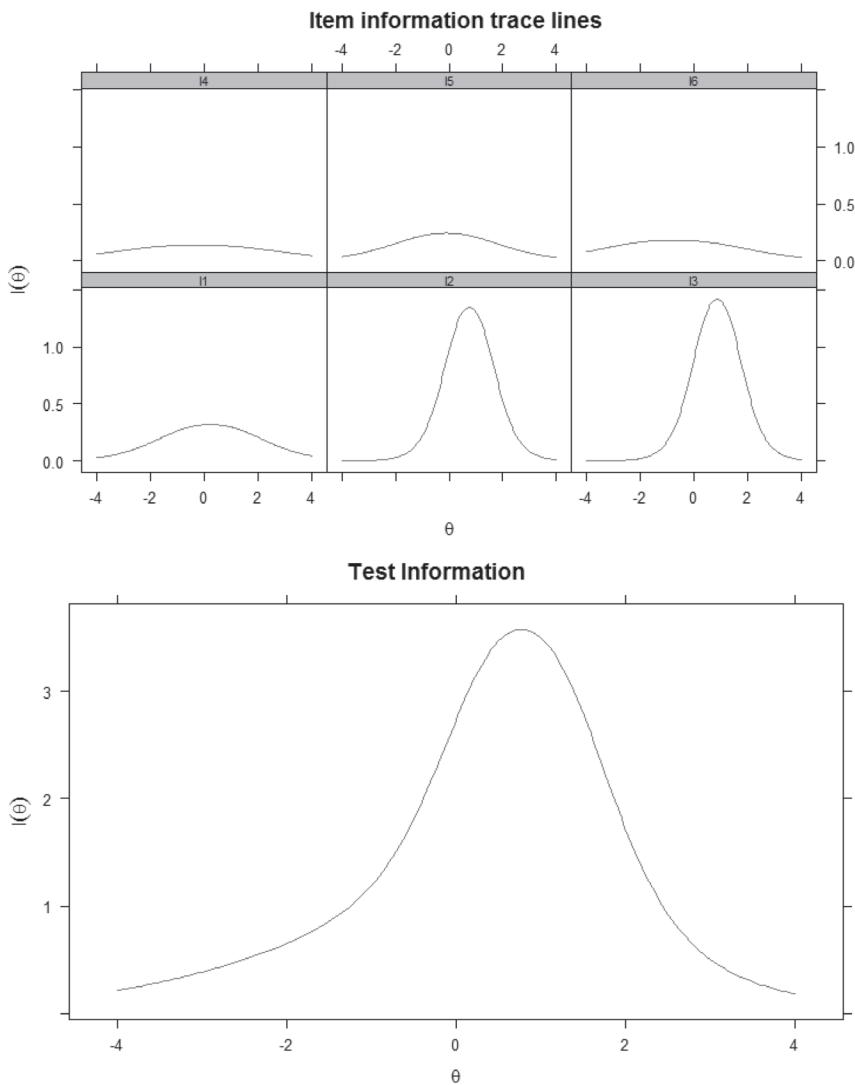


FIGURE 8.14. Item information functions (top) and total information function (bottom).

If your sex partner wants to use a condom, I'd suspect they may be having sex with someone else

strongly disagree strongly agree

Sex doesn't feel as good when you use a condom.

strongly disagree strongly agree

It's embarrassing to put on a condom (put a condom on a man).

strongly disagree strongly agree

Respondents are administered this scale and are asked to indicate on the line their degree of agreement by marking their response anywhere along the line including the

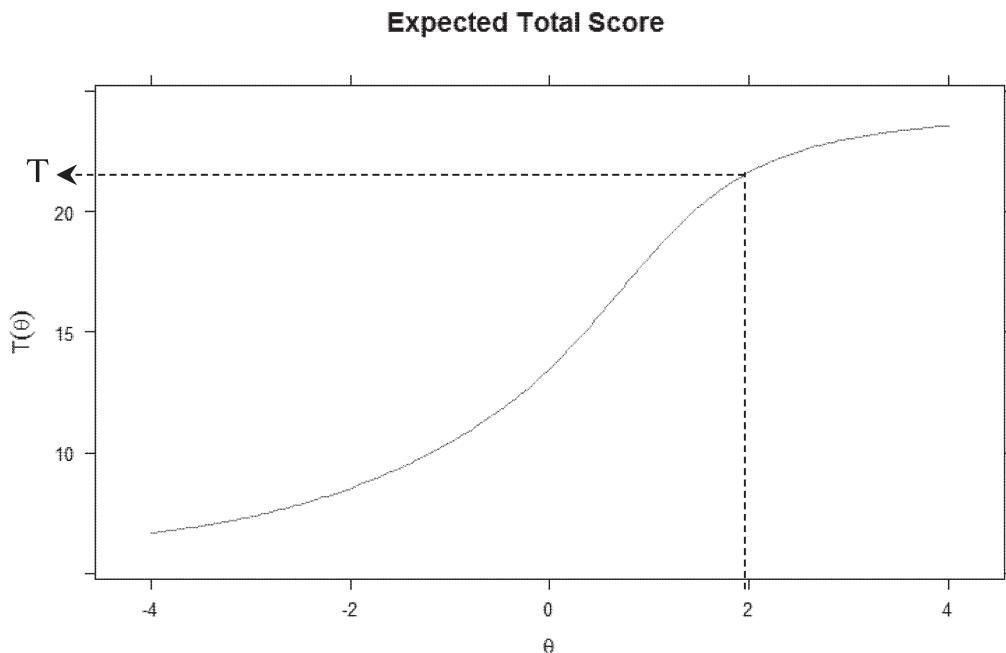


FIGURE 8.15. Expected observed score.

endpoints.⁶ If a respondent does not utilize an endpoint, then the distance from the zero point (e.g., the left end) to their response is determined. This distance can be converted to be a proportional distance. Let this proportional distance be our proportional item score, y_j . Stated another way and more generally, given a finite m_j our proportional item score y_j takes on values of $0, \frac{1}{m_j}, \frac{2}{m_j}, \dots, \frac{(m_j-1)}{m_j}, 1$ with $0 \leq y_j \leq 1$. For example, a person's responses to the above three items could be at the (left) 0-endpoint ($y_j = 0$), $\frac{3}{4}$ s of the distance from strongly disagree to strongly agree ($y_j = \frac{3}{4}$), and at the (right) 1-endpoint ($y_j = 1$). These responses would reflect a person who strongly disagreed with the partner's motivation for asking for the use of a condom, they feel that condoms really tend to adversely affect their sexual pleasure, and they strongly agree that putting a condom on a penis is embarrassing.

The psychological distance between our rating scale's endpoints ("strongly disagree," "strongly agree") can be subdivided as finely as we wish. As such, we can consider this psychological distance to be infinitely divisible and continuous in nature ($m_j = \infty$). In this case we have a *continuous item score*, \tilde{z}_j , rather than a proportional item score y_j (Samejima, 1973b).⁷ In the following we assume that respondents are not allowed to use the endpoints so that $0 < \tilde{z}_j < 1$.

From the above we have for a series of discrete graded response categories that the probability of obtaining y_j or higher is given by the GR model (Equation 8.3) or in terms of its normal ogive model representation (Samejima, 1969)

$$P_{x_j}^*(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha_j(\theta-\delta_{x_j})} e^{-t^2/2} dt. \quad (8.11)$$

By replacing the GR model's discrete x_j with the continuous item score \tilde{z}_j , we have the continuous response model (CR; Samejima, 1973b). Thus, the probability of obtaining a \tilde{z}_j or higher given θ is

$$P_{\tilde{z}_j}^*(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha_j(\theta - \tilde{\delta}_{\tilde{z}_j})} \exp(-t^2/2) dt, \quad (8.12)$$

where α_j is item j 's discrimination and $\tilde{\delta}_{\tilde{z}_j}$ is the item response difficulty parameter (Samejima, 1983). Moreover, we have the probability that a respondent selects a particular point of \tilde{z}_j is 0 as well as the limits for $P_{\tilde{z}_j}^*(\theta)$ as

$$\lim_{(\theta \rightarrow -\infty)} P_{\tilde{z}_j}^*(\theta) = 0 \quad \text{and} \quad \lim_{(\theta \rightarrow \infty)} P_{\tilde{z}_j}^*(\theta) = 1.$$

Our models for discrete item responses (e.g., $x_j = \{0, 1\}, \{0, 1, 2\}, \{1, 2, 3, 4\}$) related x_j to θ in terms of the operating characteristic function (i.e., the regression of the x_j on θ). For example, we use the term *IRF* for this operating characteristic function when x_j is dichotomous. Thus, the operating characteristic is the probability of x_j given θ (i.e., the conditional probability of x_j). With our continuous item score \tilde{z}_j the analog is called the *operating density characteristic* (i.e., conditional density distribution of \tilde{z}_j). For the CR model the operating density characteristic function (Samejima, 1973b) is

$$H_{\tilde{z}_j}(\theta) = \frac{\alpha_j}{\sqrt{2\pi}} \left[e^{-(\alpha_j^2(\theta - \tilde{\delta}_{\tilde{z}_j})^2)/2} \right] \left[\frac{d}{d\tilde{z}_j} \tilde{\delta}_{\tilde{z}_j} \right] \quad (8.13)$$

with $\int H_{\tilde{z}_j}(\theta) d\tilde{z}_j = 1.0$. This function tells us how likely \tilde{z}_j values (i.e., their distribution) are with respect to θ . For example, Figure 8.16 contains $H_{\tilde{z}_j}(\theta)$ for $\theta = -2.0$, $\theta = 0.0$, and $\theta = 1.5$ with $\alpha_j = 1.0$, $A = 1$, $B = 0$; A and B are discussed below. Focusing on $\theta = 0.0$ (dotted curve) we see that a $\tilde{z}_j = 0.5$ is more likely than are values above or below 0.5. Similarly, for $\theta = -2.0$ lower \tilde{z}_j s (e.g., $\tilde{z}_j < 0.20$) are more likely than larger values; a $\tilde{z}_j = 0.12$ is the most probable value for $\theta = -2.0$.

An alternative presentation is shown in Figure 8.17. Here we display $H_{\tilde{z}_j}(\theta)$ as a function of θ for three \tilde{z}_j s (0.20, 0.50, and 0.90) for the same α_j , A , and B as in Figure 8.16. As can be seen, if $\theta = 0$, then a response of 0.50 ($\tilde{z}_j = 0.50$) is more likely than are continuous item scores of 0.20 or 0.90 $H_{0.20j}(\theta = 0.0) = 0.15261$, $H_{0.50j}(\theta = 0.0) = 0.39894$, $H_{0.90j}(\theta = 0.0) = 0.03569$.

With the CR model $\tilde{\delta}_{\tilde{z}_j}$ is a function of \tilde{z}_j . The specific relationship between $\tilde{\delta}_{\tilde{z}_j}$ and \tilde{z}_j can be any increasing monotonic function with the mapping of \tilde{z}_j to $\tilde{\delta}_{\tilde{z}_j}$ accomplished in one of several ways. For instance, Bejar (1977) uses the relationship between an item's P_j and its location (see Appendix C, Equation C.16, to obtain $\tilde{\delta}_{\tilde{z}_j}$). Specifically and assuming \tilde{z}_j is uniformly distributed

$$\tilde{\delta}_{\tilde{z}_j} = \frac{\Phi^{-1}(\tilde{z}_j)}{r_{b_j}}, \quad (8.14)$$

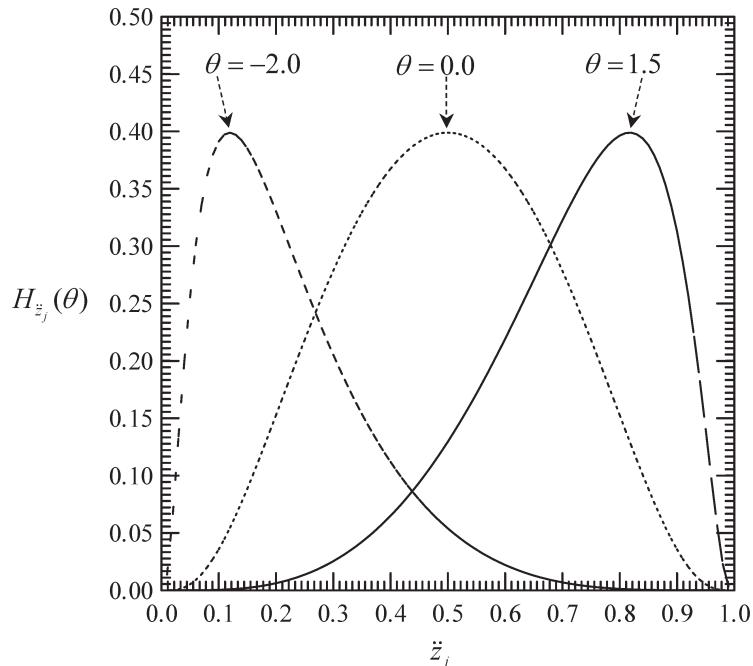


FIGURE 8.16. $H_{\ddot{z}_j}(\theta)$ s for $\theta = -2.0$, $\theta = 0.0$, and $\theta = 1.5$; $\alpha_j = 1.0$, $A = 1$, $B = 0$.

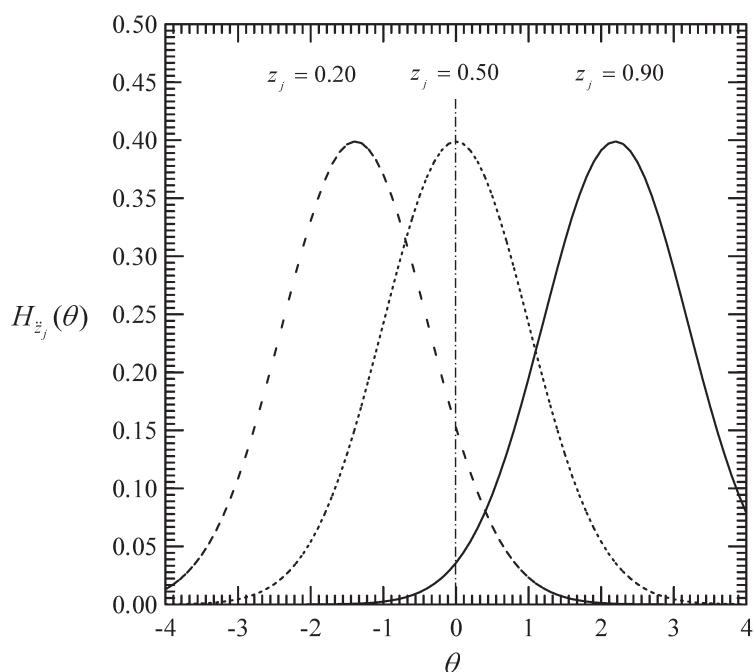


FIGURE 8.17. $H_{\ddot{z}_j}(\theta)$ s for $\ddot{z}_j = 0.20$, 0.50 , and 0.90 when $\alpha_j = 1.0$, $A = 1$, $B = 0$.

where $\Phi^{-1}(\bullet)$ is the inverse normal function and r_{bj} is our estimate of the correlation of the responses to item j and the latent variable. In contrast, in a study involving response latency \ddot{z}_j was the ratio of the time taken to the time allocated to respond to the item. In this case, Bejar (1986) used the linear function

$$\tilde{\delta}_{\ddot{z}_j} = c_j + c_j(\ddot{z}_j).$$

Samejima has used polynomials of degree K for the mapping (Samejima, 1983)

$$\tilde{\delta}_{\ddot{z}_j} = c_0 + \sum_{r=1}^K c_r \ddot{z}_j^r$$

and originally used (Samejima, 1973b)

$$\tilde{\delta}_{\ddot{z}_j} = \frac{\ln(\ddot{z}_j)}{A} - \frac{\ln(1-\ddot{z}_j)}{A} + \frac{B}{A} = \frac{\ln(\ddot{z}_j) - \ln(1-\ddot{z}_j) + B}{A}, \quad (8.15)$$

where A and B are scale and location constants, respectively.

Wang and Zeng (1998) reparameterized the CR model to have an item location on the latent continuum. Their model states the probability of a score x_j or higher as

$$P_{\ddot{z}_j}^*(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha_j(\theta-\delta_j-w_j)} \exp(-t^2/2) dt, \quad (8.16)$$

where $w_j = \ddot{z}_j/A$, $\ddot{z}_j = \ln(x_j/(m_j - x_j))$, $x_j = ((0 + \omega), \dots, (m_j - \omega))$, and ω is the unit of measurement; Wang and Zeng state that m_j does not have to be an integer. In this formulation δ_j and x/m_j correspond to Samejima's B/A and \ddot{z}_j , respectively.

As an example of the CR model's cumulative probability trace lines, assume that our continuous rating scale is 5 cm long ($m_j = 5$) and we measure a person's response from the left endpoint. Assuming a unit of measurement of ($\omega = 0.1$), there are 49 cumulative probability trace lines in the open response situation. We select three x_j 's from this set: 2.3 cm, 3.1 cm, and 4.7 cm. To use Equations 8.12, these x_j 's are converted to proportions ($\ddot{z}_j = x/m_j$). Thus, the \ddot{z}_j 's corresponding to 2.3 cm, 3.1 cm, and 4.7 cm are 2.3/5 = 0.46, 0.62, and 0.94, respectively. Alternatively, we can forego calculating the \ddot{z}_j 's and directly use the x_j 's in Equation 8.16. Figure 8.18 shows the cumulative probability trace lines when $\alpha_j = 1.5$, $B = 0.5$, and three A levels (1, 2, 10).⁹ As can be seen, as A increases, the spacing between the trace lines decreases; B 's value simply shifts the trace line set up or down the continuum. As is the case with the models from the preceding chapters, the item discrimination, α_j , affects the slope of the corresponding trace line.

Analogous to the discrete score polytomous models, each continuous item score provides information for estimating $\hat{\theta}$ (Samejima, 1973b)

$$I_{\ddot{z}_j}(\theta) = \alpha_j^2, \quad (8.17)$$

Equation 8.17 shows that the amount of information provided by a score is the same

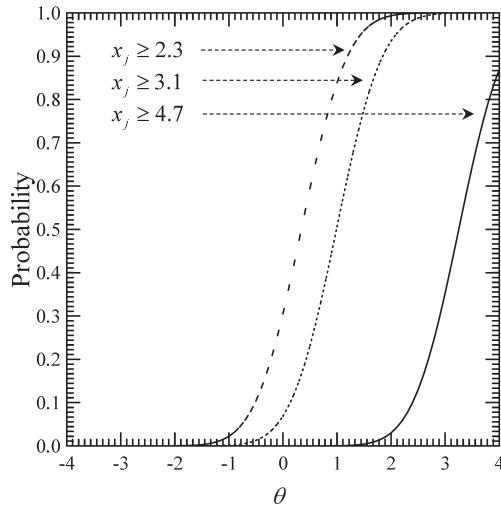
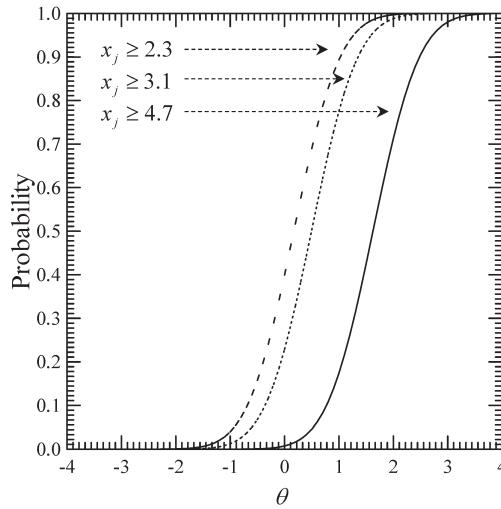
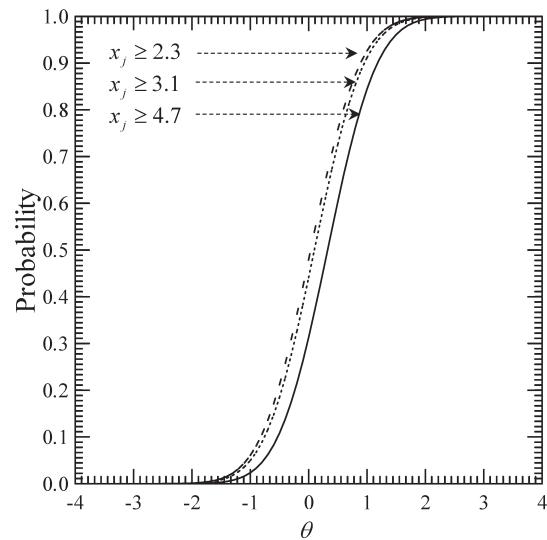
$A = 1.0$  $A = 2.0$  $A = 10.0$ 

FIGURE 8.18. Cumulative probability trace lines for $x_j = (2.3, 3.1, 4.7)$, $\alpha_j = 1.5$, $A = 1, 2, 10$, and $B = 0.5$.

for θ and \ddot{z}_j . The total information is the sum of the continuous score information functions

$$I(\theta) = \sum_j^L I_{\ddot{z}_j}(\theta) = \sum_j^L \alpha_j^2, \quad (8.18)$$

Parameter estimation may be accomplished in different ways. For example, Same-

jima (1973b) discusses a factor analytic-based approach to estimate item discrimination and obtain the $\hat{\delta}_{ij}$ s (also see Appendix C, Bejar, 1977, and Ferrando, 2002). Shojima (2005) presents both JMLE and MMLE, along with MLE and EAP person location estimation. The MMLE approach is implemented in the R package EstCRM (Zopluoglu, 2012, 2015). More information on the CR model's estimation may be found in Zopluoglu (2013).

Summary

The generalized partial credit model relaxes the equal item discrimination assumption of the partial credit model. Thus, the GPC model contains a discrimination parameter that indicates the degree to which an item can differentiate among different values of θ . As in the partial credit model, the points of intersection of adjacent ORFs are the transition location parameters, δ_{jh} s. These transition locations do not need to be sequentially ordered. The model's ORFs reflect an interaction between the item's transition location parameters and discrimination parameter. The GPC may be applied to situations where ordered response data arise through either partial credit scoring, ratings, or the use of a Likert response format. In this latter case, the GPC may be referred to as the generalized rating scale model, that is, a model that has a set of transition location parameters that are constant across items, but varying item discrimination. Therefore, the GRS model allows for the modeling of Likert response data where items are allowed to vary in their discrimination.

The partial credit model is (mathematically) a special case of the generalized partial credit model in the same way that the Rasch model is a special case of the 2PL model. However, some individuals consider the partial credit model to represent a philosophical approach to measurement that is not reflected in the generalized partial credit. (This philosophical difference is the same one presented with the Rasch model in Chapter 2.) The GPC model simplifies to the 2PL model in the two-category (dichotomous) case. As is the case with the partial credit and rating scale models, the generalized partial credit model is a special case of the nominal response model.

Another model for ordered polytomous data that allows items to vary in their discrimination is the graded response model. Both Masters (1982) and Muraki's (1992) original formulations of their models are in terms of a series of dichotomous models that govern the probability of responding in one category versus the next adjacent category. In contrast, the formulation of the graded response model involves specifying the probability that a respondent would obtain a given category score or higher versus one or more lower category scores. As such, the model provides the cumulative probabilities of obtaining different subsets of category scores. To obtain the probability of responding in a particular category, one subtracts the cumulative probabilities for adjacent scores from one another. Unlike the case with the partial credit, rating scale, and generalized partial credit models, the category boundary location parameters are sequentially ordered in the graded response model. Given this characteristic, we need to examine the ORFs to determine whether each response category is the most likely response at some

point along the continuum. The GR model simplifies to the two-parameter model when applied to dichotomous data. It should be noted that in the context of proficiency assessment, neither the GPC model nor the GR model addresses the possibility of examinees guessing on items. The GR model can be considered to be a special case of the continuous response model. The CR model replaces the GR model's discrete item score x_j with a continuous item score \tilde{z}_j . As such, the CR model is appropriate for use with items that utilized a continuous response scale or that use time as a person measure.

In the next chapter we present a model for polytomous responses that are not inherently ordered. Such data may arise, for example, in proficiency testing (e.g., analogy items using a multiple-choice item format) or with attitudinal or survey instruments that use a nonordered response format (e.g., "Yes," "No," "Unsure"). This model, the nominal response model, contains two parameters for each response category to reflect the attractiveness of each category as well as how well each category differentiates among θ s. As mentioned above, the nominal response model can be constrained to subsume the GPC, GRS, PC, and RS models as well as the two-parameter model.

Notes

1. The models are, in essence, *ordinal logistic regression* models using latent person and item characterizations. In the case of the graded response model, a probit link function is used in lieu of the logit link function. That is, the function is the cumulative density function of the normal distribution. Therefore, we have $\text{probit}[p(b)] = \gamma + \alpha\theta$ rather than $\text{logit}[p(b)] = \gamma + \alpha\theta$.
2. Samjima's (1969) development is based on the two-parameter normal ogive model (Equation 8.22; also see Appendix C). As such, the presentation of the GR logistic model typically includes the scaling constant D

$$P_{x_j}^*(\theta) = \frac{e^{D\alpha_j(\theta - \delta_{x_j})}}{1 + e^{D\alpha_j(\theta - \delta_{x_j})}} \quad (8.19)$$

However, the perspective in this book is that the logistic metric is intrinsically useful, and we are not concerned with approximating the normal metric. Therefore, the GR model in Equation 8.3 does not include D .

3. When the GR model is applied to dichotomous data, the GR model simplifies to the two-parameter model. When an item has only two possible scores (e.g., $x_j = 0$ or 1), then $m_j = 1$ and the probability of a response of 1, p_1 , equals

$$p_1 = P_1^* - P_2^* = P_1^* - 0 = \frac{e^{\alpha_j(\theta - \delta_1)}}{1 + e^{\alpha_j(\theta - \delta_1)}} - 0 = \frac{e^{\alpha_j(\theta - \delta)}}{1 + e^{\alpha_j(\theta - \delta)}} \quad (8.20)$$

because $P_2^* \equiv 0$ and the probability of a response of 0 is

$$p_0 = P_0^* - P_1^* = 1.0 - P_1^* = 1.0 - \frac{e^{\alpha_j(\theta - \delta_1)}}{1 + e^{\alpha_j(\theta - \delta_1)}} = \frac{1}{1 + e^{-\alpha_j(\theta - \delta)}} \quad (8.21)$$

because $P_0^* \equiv 1.0$. Equation 8.20 is the 2PL model and shows that the GR model is equivalent to the two-parameter model when an item has only two response categories; with only two categories, there is only one δ and the category subscript on δ is dropped. If discrimination is held constant across items, then Equation 8.20 becomes the one-parameter model.

- As an example of the calculations involved in obtaining Figures 8.8 and 8.9, assume $\alpha = 1.5$, $\delta_{1j} = -1.0$, $\delta_{2j} = 1.0$, and $\theta = 0.5$. Therefore, the probability of responding in category 0 ($x_j = 0$) is

$$\begin{aligned} p_0 = P_0^* - P_1^* &= p(x_j = \{0, 1, 2\} | \theta) - p(x_j = \{1, 2\} | \theta) = 1.0 - \frac{e^{\alpha_j(\theta - \delta_{1j})}}{1 + e^{\alpha_j(\theta - \delta_{1j})}} \\ &= 1.0 - \frac{e^{1.5(0.5 - (-1.0))}}{1 + e^{1.5(0.5 - (-1.0))}} = 1.0 - 0.9047 = 0.0953 \end{aligned}$$

the probability of responding in category 1 ($x_j = 1$) is

$$\begin{aligned} p_1 = P_1^* - P_2^* &= p(x_j = \{1, 2\} | \theta) - p(x_j = 2 | \theta) = \frac{e^{\alpha_j(\theta - \delta_{1j})}}{1 + e^{\alpha_j(\theta - \delta_{1j})}} - \frac{e^{\alpha_j(\theta - \delta_{2j})}}{1 + e^{\alpha_j(\theta - \delta_{2j})}} \\ &= \frac{e^{1.5(0.5 - (-1.0))}}{1 + e^{1.5(0.5 - (-1.0))}} - \frac{e^{1.5(0.5 - 1.0)}}{1 + e^{1.5(0.5 - 1.0)}} = 0.9047 - 0.3208 = 0.5838 \end{aligned}$$

and the probability of responding in category 2 ($x_j = 2$) equals

$$\begin{aligned} p_2 = P_2^* - P_3^* &= p(x_j = 2 | \theta) - p(x_j > 2 | \theta) = \frac{e^{\alpha_j(\theta - \delta_{2j})}}{1 + e^{\alpha_j(\theta - \delta_{2j})}} - 0 = P_2^* \\ &= \frac{e^{1.5(0.5 - 1.0)}}{1 + e^{1.5(0.5 - 1.0)}} - 0 = 0.3208 - 0 = 0.3208. \end{aligned}$$

The sum of these conditional probabilities is 1.0.

To obtain the corresponding probabilities with the normal ogive version of the GR model

$$\pi_j^*(x_j = 1) = \int_{-\infty}^{\alpha_j(\theta - \delta_{x_j})} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2)}{2}\right) dz \quad (8.22)$$

We use the Excel function =NORM.DIST($(\alpha_j(\theta - \delta_{x_j}))$, 0, 1, TRUE) to obtain our probabilities

$$\pi_0^*(x_j = 0, 1, 2) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2)}{2}\right) dz = 1.0$$

$$\pi_1^*(x_j = 1, 2) = \int_{-\infty}^{1.5(0.5 - (-1))} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2)}{2}\right) dz = 0.9878$$

$$\pi_2^*(x_j = 2) = \int_{-\infty}^{1.5(0.5-1)} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2)}{2}\right) dz = 0.2266$$

$$\pi_3^*(x_j > 2) = \int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2)}{2}\right) dz = 0.0$$

Therefore, the probability of responding in category 0 ($x_j = 0$) is

$$p_0 = P_0^* - P_1^* = \pi_0^*(x_j = 0, 1, 2) - \pi_1^*(x_j = 1, 2) = 1.0 - 0.9878 = 0.0122$$

the probability of responding in category 1 ($x_j = 1$) is

$$\begin{aligned} p_1 &= P_1^* - P_2^* = \pi_1^*(x_j = 1, 2) - \pi_2^*(x_j = 2) \\ &= \int_{1.5(0.5-1)}^{1.5(0.5-(-1))} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2)}{2}\right) dz = 0.9878 - 0.2266 = 0.7611 \end{aligned}$$

and the probability of responding in category 2 ($x_j = 2$) equals

$$p_2 = P_2^* - P_3^* = \pi_2^*(x_j = 2) - \pi_3^*(x_j > 2) = 0.2266 - 0 = \pi_2^*(x_j = 2) = 0.2266.$$

The sum of these conditional probabilities is 1.0. The differences in probabilities using Equations 8.3 and 8.22 are due to not using the scaling constant D (i.e., not using Equation 8.19).

5. A graphic response scale typically has descriptors along the line. For example two graphic response scales from the Graphic Rating Report on Workers (Patterson, 1922) are

Very Superior	Learns with Ease	Ordinary	Slow to Learn	Dull
Usually High Output	Satisfactory Output	Limited Output	Unsatisfactory Output	

The term *visual analog scale* has been used in lieu of graphic response scale. We will use the term *continuous rating scale* because we will not use descriptors along the line. In implementations these continuous rating scales may be called a slider scale or a sliding response scale.

6. Samejima (1973b) distinguishes between an “open response situation” or a “close response situation” depending on whether the respondent is allowed to use the end-

points. In an “open response situation” the respondent cannot use the endpoints, whereas in a “close response situation” the respondent is allowed to use the endpoints. In short, with an “open response situation” the probabilities assigned to the endpoints are zero, whereas in the “close response situation” the probabilities assigned to the endpoints are nonzero.

7. For an “open response situation” we have that $0 < \ddot{z}_j < 1$ and for a “close response situation” we have that $0 \leq \ddot{z}_j \leq 1$.
8. The logistic version of Equation 8.12 is

$$P_{\ddot{z}_j}^*(\theta) = \frac{e^{\alpha_j(\theta - \tilde{\delta}_{\ddot{z}_j})}}{1 + e^{\alpha_j(\theta - \tilde{\delta}_{\ddot{z}_j})}} \quad (8.23)$$

and the operating density characteristic function

$$H_{\ddot{z}_j}(\theta) = \alpha_j P_{\ddot{z}_j}^*(\theta) \left[1 - P_{\ddot{z}_j}^*(\theta) \right] \left[\frac{d}{d\ddot{z}_j} \tilde{\delta}_{\ddot{z}_j} \right] \quad (8.24)$$

9. These As (1, 2, 10) and B (0.5) correspond to Equation 8.16’s δ_j values of 0.5, 0.25, 0.05, respectively.

9

Models for Nominal Polytomous Data

In this chapter we discuss models for nominal polytomous data. Unlike ordered polytomous data, with nominal data the response categories are not inherently ordered. As a result, one does not have a direct, or an inverse, relationship between the observed responses and the magnitude of θ . However, it is still possible to capture information in each of the possible responses. Therefore, in addition to the previously mentioned advantages of IRT over classical test theory (CTT), these models provide a person location estimate that is directly based on the individual's nominal response data. In contrast, because the sum of nominal polytomous responses has no inherent meaning vis-à-vis the variable being measured, the traditional approach of using a total score as a person measure would be meaningless. The best that one can do in the traditional sense is to sum the *dichotomized* polytomous responses to obtain a potentially meaningful total score.

Nominal response data (also known as categorical response data) may arise in a number of situations. For instance, in the social sciences scales will sometimes incorporate variables that use polytomous categorical response formats. Responses to these variables constitute nominal response data. For example, consider the following three-item scale for measuring compulsive behavior using a categorical response format.

1. Whenever I leave the house I feel a need to check at least five times that I locked each door and window.
 Yes No Maybe Won't Say
2. I need to wash my hands at least five times before I can eat.
 Yes No Maybe Won't Say
3. Whenever I mail a check to pay a bill I will verify that I have signed the check at least five times before mailing it.
 Yes No Maybe Won't Say

None of the previously discussed models would be appropriate for modeling all four categorical responses to these items.

In other situations, we might specifically design questions to incorporate useful information for locating an individual. For instance, research on student misconceptions in solving mathematics problems (e.g., Brown & Burton, 1978; Brown & VanLehn, 1980; Tatsuoka, 1983) has shown that incorrect responses can be due to more than just one kind of misconception. As such, items may be designed to incorporate these misconceptions rather than simply providing arbitrary or plausible incorrect alternatives. However, it may not be possible to order the item alternatives to reflect different degrees or severity of misconceptions. In this case one would have unordered categorical response data. As an example of an item constructed to incorporate misconceptions, consider the following item whose alternatives reflect the application of erroneous rules of signed-number subtraction (Tatsuoka, 1983):

1. $-6 - (-10) = ?$
 - a. -16
 - b. -4
 - c. 4

These alternatives potentially provide useful information, not only for locating individuals, but also for the diagnosing mathematical misconceptions. If we dichotomized the responses into incorrect and correct categories, then we would be discarding this information. Because the incorrect alternatives do not represent partially correct answers, the use of models for ordered polytomous data (e.g., the PC, GPC, or GR models) would be inappropriate for modeling this item and others like it.

Conceptual Development of the Nominal Response Model

Polytomous categorical data consist of mutually exclusive unordered response categories. These response categories may correspond to (1) item options from a multiple-choice item format, (2) response options from a survey or attitude instrument, and (3) rater judgments (rater judgments do not necessarily result in ordered ratings), to name just a few possibilities. Figure 9.1 contains a graphic depiction of the response format for the above compulsive behavior scale item, "I need to wash my hands at least five times before I can eat." In the figure the ellipse represents the item, and the response categories are presented as not being contiguous with one another to avoid implying that they are inherently ordered.

Conceptually, each of the item's response categories has an associated probability such that the sum of these response probabilities across the categories is 1.¹ For example, when this item is administered to an infinitely large sample, the relative frequency of individuals responding "yes," "no," "maybe," and "won't say" might be, for example, 0.50, 0.25, 0.20, and 0.05, respectively. Let the probabilities associated with the responses "yes," "no," "maybe," and "won't say" be $p_1 = 0.50$, $p_2 = 0.25$, $p_3 = 0.20$, and

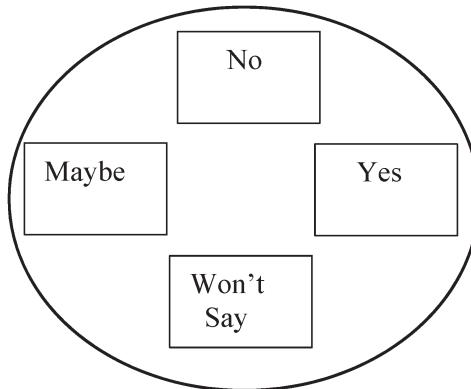


FIGURE 9.1. Schematic of nominal response categories.

$p_4 = 0.05$. Based on this set of probabilities, we can determine the odds of one response versus another. For example, the odds of a response being a “no” as opposed to a “yes” are $odds_{2,1} = (p_2 / p_1) = 1/2$. Thus, a response of “no” is half as likely as a response of “yes,” or we expect to see one “no” for every two “yes’s.” (Appendix G, “Odds, Odds Ratios, and Logits,” contains an introduction to odds.)

As mentioned in Chapter 2, for convenience odds can be transformed to a logarithmic scale to yield the log odds (i.e., the logit). As an example, given that the response is either a “no” or a “yes,” then the log odds of a response being a “no” as opposed to a “yes” are $\log(p_2 / p_1)$. Moreover, this logit transformation allows one to express the log odds of a response being a “no” as opposed to a “yes” in terms of a predictor X (e.g., an individual’s compulsive behavior)

$$\log(p_2 / p_1) = \gamma_2^u + \alpha_2^u X,$$

where α_2^u and γ_2^u characterize the “no” response category. (The term “ $\gamma + \alpha X$ ” is the slope–intercept parameterization form of the logit seen in Chapter 2.) The symbol γ^u is the intercept and reflects the propensity to respond in one category over the other category regardless of X (i.e., γ^u is the baseline log odds—a model without a predictor). In addition, α^u is the change in the log odds, or the logit, for this response as the predictor X changes by one unit, that is, the slope of the logit regression line. If $\alpha^u = 0$, then the log odds do not change as X changes.

Similarly, the log odds of a response being a “maybe” as opposed to a “yes” may be obtained by

$$\log(p_3 / p_1) = \gamma_3^u + \alpha_3^u X$$

and the log odds of a response being a “won’t say” as opposed to a “yes” is

$$\log(p_4 / p_1) = \gamma_4^u + \alpha_4^u X$$

The above three logit equations use the “yes” response category as a *baseline* response category (i.e., the odds or log odds are with respect to the baseline category). This category may be one of specific interest or may be the category that has the largest frequency (Agresti, 1990).

Rather than talking about the log odds of a response in one category over another category, the probability of a particular response can be directly expressed by using the above logits.² Using these logit equations, we have the conditional probability for each response category, given X, as follows:

$$\begin{aligned} p(x=1|X) &= \frac{e^{\gamma_1^u + \alpha_1^u X}}{e^{\gamma_1^u + \alpha_1^u X} + e^{\gamma_2^u + \alpha_2^u X} + e^{\gamma_3^u + \alpha_3^u X} + e^{\gamma_4^u + \alpha_4^u X}}, \\ p(x=2|X) &= \frac{e^{\gamma_2^u + \alpha_2^u X}}{e^{\gamma_1^u + \alpha_1^u X} + e^{\gamma_2^u + \alpha_2^u X} + e^{\gamma_3^u + \alpha_3^u X} + e^{\gamma_4^u + \alpha_4^u X}}, \\ p(x=3|X) &= \frac{e^{\gamma_3^u + \alpha_3^u X}}{e^{\gamma_1^u + \alpha_1^u X} + e^{\gamma_2^u + \alpha_2^u X} + e^{\gamma_3^u + \alpha_3^u X} + e^{\gamma_4^u + \alpha_4^u X}}, \end{aligned}$$

and

$$p(x=4|X) = \frac{e^{\gamma_4^u + \alpha_4^u X}}{e^{\gamma_1^u + \alpha_1^u X} + e^{\gamma_2^u + \alpha_2^u X} + e^{\gamma_3^u + \alpha_3^u X} + e^{\gamma_4^u + \alpha_4^u X}}.$$

More generally, letting m represent the number of response categories, we have the probability of a response, k , conditional on X is

$$p(x=k|X) = \frac{e^{\gamma_k^u + \alpha_k^u X}}{\sum_{h=1}^m e^{\gamma_h^u + \alpha_h^u X}} = \frac{e^{\gamma_k^u + \alpha_k^u X}}{1 + \sum_{h=2}^m e^{\gamma_h^u + \alpha_h^u X}}. \quad (9.1)$$

To identify this model we set the baseline response category’s α^u and γ^u to 0. For instance, using the first response category as the baseline response category, we have $\alpha_1^u = \gamma_1^u = 0$. As a result, there are $m - 1$ unique logit equations with $(m - 1) \alpha_k^u$ and $(m - 1) \gamma_k^u$ s. In other words, there are $(m - 1)$ potentially nonzero α_k^u s, $(m - 1)$ potentially nonzero γ_k^u s as well as one α_k^u and γ_k^u that are each equal to zero. The probability for the baseline category may be obtained by subtracting the sum of the other categories’ probabilities from 1.

The $(m - 1) \alpha_k^u$ s may be transformed to obtain each of the m response categories’ α_k . Similarly, the $(m - 1) \gamma_k^u$ s may be transformed so that each of the $m \gamma_k$ s is potentially nonzero. (Note that we use the superscript “u” to distinguish between the estimated slopes and intercepts [i.e., the $(m - 1) \alpha_k^u$ s and $(m - 1) \gamma_k^u$ s] and the transformed slopes and intercepts [i.e., the $m \alpha_k$ s and $m \gamma_k$ s].) With these transformations, the $\alpha_1^u = \gamma_1^u = 0$ constraint will “translate” into the $m \alpha_k$ s collectively summing to zero, and the $m \gamma_k$ s, as a set, will be constrained to sum to zero (i.e., the model is still identified). These transformations are accomplished by using an appropriate transformation matrix.³

As an example of these transformations, assume a four-response category item ($m = 4$) with α'_k^u s of $\{0.44, -0.55, -0.60\}$ and γ'_k^u s of $\{0.37, -0.74, -0.34\}$. (The prime symbol on $\underline{\alpha}^u$ and $\underline{\gamma}^u$ indicates that these are row vectors, not column vectors.) If we use the transformation matrix, \mathbf{T} , with the values of

$$\mathbf{T} = \begin{bmatrix} -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{bmatrix}$$

we obtain (through matrix algebra)

$$\alpha_1 = 0.44*(-0.25) + (-0.55*-0.25) + (-0.60*-0.25) = 0.18$$

$$\alpha_2 = 0.44*(0.75) + (-0.55*-0.25) + (-0.60*-0.25) = 0.62$$

$$\alpha_3 = 0.44*(-0.25) + (-0.55*0.75) + (-0.60*-0.25) = -0.37$$

$$\alpha_4 = 0.44*(-0.25) + (-0.55*-0.25) + (-0.60*0.75) = -0.42$$

The sum of these α_k s is 0.0. For the intercepts we have

$$\gamma_1 = 0.37*(-0.25) + (-0.74*-0.25) + (-0.34*-0.25) = 0.18$$

$$\gamma_2 = 0.37*(0.75) + (-0.74*-0.25) + (-0.34*-0.25) = 0.55$$

$$\gamma_3 = 0.37*(-0.25) + (-0.74*0.75) + (-0.34*-0.25) = -0.56$$

$$\gamma_4 = 0.37*(-0.25) + (-0.74*-0.25) + (-0.34*0.75) = -0.16$$

These γ_1 s also sum to 0; the α_k s and γ_k s reflect the transformed α'_k^u s and γ'_k^u s. With the α_k s and γ_k s we potentially have a nonzero slope–intercept pair for each response category.

Equation 9.1 can be cast into the context of a latent person location variable and latent item parameters. The result is Bock's (1972) *nominal response* (NR) model (also called the *nominal categories* model). According to the NR model, the probability of a person located at θ responding in item j 's k th category is

$$p_j(x=k|\theta, \underline{\alpha}, \underline{\gamma}) = \frac{e^{\gamma_{jk} + \alpha_{jk}\theta}}{\sum_{h=1}^{m_j} e^{\gamma_{jh} + \alpha_{jh}\theta}}, \quad (9.2)$$

where α_{jk} and γ_{jk} are the slope and the intercept parameters, respectively, of the response function associated with the k th indexed category of item j , and m_j is the number of response categories of item j (i.e., $k = \{1, \dots, m_j\}$). (Note the italicized “ m ” indicates the number of response categories, not the number of transition locations as seen with, say, the PC model.) The symbol k simply indexes the response categories and does not imply

that the categories are ordered. Unlike the model in Equation 9.1, the NR model is in terms of the *transformed* α_{jk} s and γ_{jk} s. For item j , γ_{jk} reflects the individual's propensity to use response category k , and α_{jk} reflects, in part, the option's discrimination capacity. The magnitude of the α_{jk} s also reflects the "order" of the categories because of the constraint mentioned above and discussed below. For convenience the m_j category slope and intercept parameters are sometimes collected into vectors: $\underline{\alpha}' = (\alpha_{j1}, \dots, \alpha_{jm})$ and $\underline{\gamma}' = (\gamma_{j1}, \dots, \gamma_{jm})$. In the following, for brevity we use p_{jk} in lieu of $p_j(x = k | \theta, \underline{\alpha}, \underline{\gamma})$.

The indeterminacy issue previously discussed may be addressed in one of two ways. One approach is to constrain each parameter type to sum to zero ($\sum_{k=1}^{m_j} \alpha_{jk} = 0$, $\sum_{k=1}^{m_j} \gamma_{jk} = 0$). Alternatively, the α and γ for a baseline response category may be set to 0. As a result of these constraints, the number of estimated category slopes and intercepts for item j is $2(m_j - 1)$.

As discussed in connection with Equation 9.1, these constraints are implemented by using a transformation matrix, \underline{T} , on the *unconstrained* slope, α_{jk}^u , and *unconstrained* intercept, γ_{jk}^u , parameters. By unconstrained, we mean the slope and intercept parameters prior to imposition of the constraints by the transformed matrix \underline{T} . These unconstrained discrimination and intercept parameters can be collected into vectors $\underline{\alpha}^u = (\alpha_{j1}^u, \dots, \alpha_{j,m_j-1}^u)$ and $\underline{\gamma}^u = (\gamma_{j1}^u, \dots, \gamma_{j,m_j-1}^u)$.

To obtain the *constrained* set of slopes ($\underline{\alpha}$) and intercepts ($\underline{\gamma}$), one postmultiplies the corresponding estimated unconstrained parameter estimate vectors by \underline{T}

$$\underline{\alpha} = \underline{\alpha}^u \underline{T}$$

and

$$\underline{\gamma} = \underline{\gamma}^u \underline{T}$$

Thus, we can obtain the constrained slope and intercept parameter estimates from the estimated unconstrained slope and intercept parameters. As a result, for each item we have m_j constrained α_{jk} s and γ_{jk} s parameters where the m th discrimination parameter, α_{jm} , is equal to $1 - \sum_{k=1}^{m_j-1} \alpha_{jk}$ and the m th intercept parameter, γ_{jm} , is $1 - \sum_{k=1}^{m_j-1} \gamma_{jk}$ for $k = \{1, \dots, m_j\}$. See Bock (1972) as well as Thissen et al. (2003) for greater detail.

Equation 9.2 is sometimes written in a multivariate logit (Bock, 1972) or multinomial logit (Bock, 1997) form

$$p_{jk} = \frac{e^{z_{jk}(\theta)}}{\sum_{h=1}^{m_j} e^{z_{jh}(\theta)}}, \quad (9.3)$$

where $z_{jk}(\theta) = \gamma_{jk} + \alpha_{jk}\theta$ and is the multivariate or multinomial logit. In this form the constraints of $\sum_h \alpha_{jh} = 0$ and $\sum_h \gamma_{jh} = 0$ become $\sum_h z_{jh}(\theta) = 0$ for $h = \{1, \dots, m_j\}$.

To aid in interpreting these parameters, we present a logit space plot of the logit (i.e., $\gamma_{jk} + \alpha_{jk}\theta$) as a function of θ for a three-category ($m_j = 3$) item in Figure 9.2; $\underline{\alpha}' = (-0.75, -0.25, 1.0)$ and $\underline{\gamma}' = (-1.5, -0.25, 1.75)$. Recall from Chapter 2 that the lines in this graph are logit regression lines. As can be seen, the γ_{jk} value is the y-intercept (i.e.,

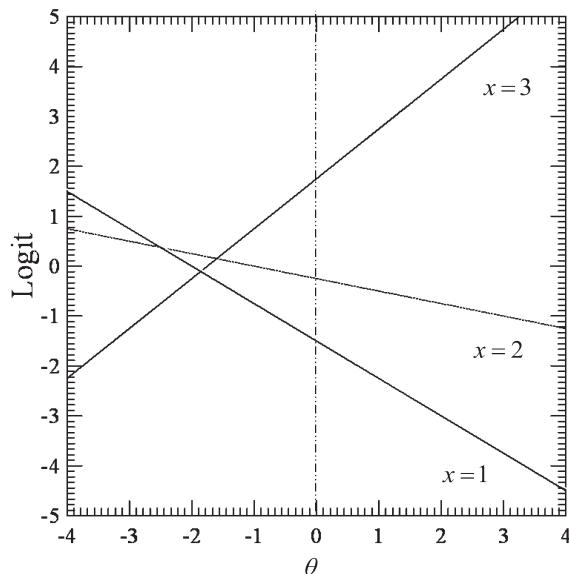


FIGURE 9.2. NR model logit regression lines for a three-category item with $\alpha' = (-0.75, -0.25, 1.0)$ and $\gamma' = (-1.5, -0.25, 1.75)$.

when $\theta = 0.0$) and α_{jk} is the slope of the corresponding logit regression line and indicates how the log odds for response category k change as the individual's θ increases. Because category 3 has the largest α_{jk} , its corresponding logit regression line has the steepest positive slope. Conversely, the largest negative α_{jk} is associated with category 1, and its logit regression line has the maximal negative slope. Of the three options, category 3 differentiates the best. For this category an increase from $\theta = 0$ to 1 corresponds to an increase in the log odds for this category from 1.75 to 2.75 (i.e., the difference is the value of α_3). Furthermore, we see that the log odds of persons located in the upper end of the continuum favor selecting category 3 over the other response categories. Conversely, it can be seen that individuals located at the lowest end of the scale will tend to select category 1 over categories 2 and 3, although there is a strong tendency for these individuals to also select category 2. Furthermore, as an individual's location increases, the tendency to select either categories 1 and 2 decreases.

In general, Figure 9.2 is somewhat typical of the logit space plot for an item. As would be expected, if the α_{jk} s for two or more response categories are equal and these categories have unequal γ_{jk} s, then the corresponding logit regression lines are parallel. In contrast, if the α_{jk} s for two or more response categories are equal and have equal γ_{jk} s, then the response categories collapse into a single category.

Although Figure 9.2 tells us the log odds of selecting a category as a function of θ , it is sometimes difficult to look at a logit regression line and get a sense of the probability of responding in the corresponding category. For this purpose we can examine the item's ORFs. Figure 9.3 contains the ORFs corresponding to the item shown in Figure 9.2; these ORFs are obtained by using Equation 9.2.

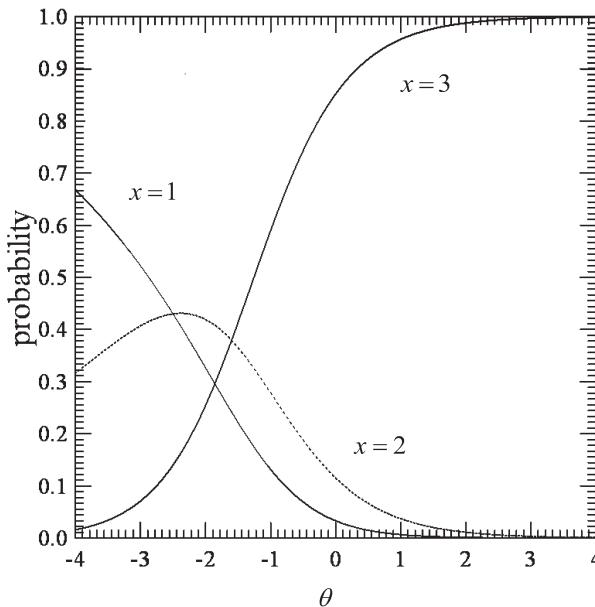


FIGURE 9.3. NR model ORFs for the three-category item shown in Figure 9.2.

A comparison of Figures 9.2 and 9.3 shows that the points of intersection of the logit regression lines correspond to the transition points among the ORFs. In general, with the NR model an item has one category ORF that has a maximal positive slope (monotonically increasing) and one that has a maximal negative slope (monotonically decreasing). The former is typically the correct response and/or the category with the highest frequency. The other response category(ies) have usually, but not always, a unimodal pattern. However, the overall pattern of an item's ORFs as well as their points of intersection reflect an interaction among the item's α_{jk} s and γ_{jk} s. In addition, if the α_{jk} s for two or more response categories are equal and have unequal γ_{jk} s, then the corresponding ORFs do not intersect, and the response category with the larger γ_{jk} envelops the other(s). In this case, the response categories discriminate the same as one another but are differentially attractive. Unless there is a substantive reason to maintain separate response categories, one might consider combining these response categories. When response categories have equal α_{jk} s as well as equal γ_{jk} s, then these response categories have collapsed into a single category with a single ORF.

We can formulate category location parameters in two ways. The first approach is analogous to that of the PC model (Chapter 7) in which the location parameters are defined at the intersection of ORFs. However, this does not mean that we are assuming the response categories are ordered. The ORFs' intersection point can be obtained by setting the corresponding category multivariate logits equal to one another and solving for θ . Therefore, after simplification, noting that θ and δ are on the same scale, and dropping the item subscript for convenience, we have that for any item with $m_j \geq 2$, $k^* < k$, and $\alpha_{k^*} \neq \alpha_k$ the transition point between categories k^* and k is

$$\delta_{k^*,k} = \frac{\gamma_{k^*} - \gamma_k}{\alpha_k - \alpha_{k^*}} \quad (9.4)$$

As is true for the PC and GPC models, with the NR model there are $m_j - 1$ transition location parameters across categories. If $\alpha_{k^*} = \alpha_k$, then the k^* th ORF does not intersect with the k th ORF and $\delta_{k^*,k}$ is undefined. For a binary item (i.e., $m_j = 2$), the 2PL and NR models are equivalent and the transition point is given by δ . Therefore, the NR model can model both dichotomous and polytomous data. As mentioned in the preceding chapters, the PC, RS, GPC, and GRS models are special cases of the NR model. Thissen and Steinberg (1986) show that when the α_{jk} s are forced to increase in steps of one, the NR model becomes the PC/GPC models; also see Sijtsma and Hemker (2000) and Muraki (1992). As such, the NR model may also be applied to ordinal data. In fact, Mellenbergh (1995) shows different ways of preserving the order of the response categories for use with the NR model. These methods yield different types of models (e.g., adjacent-category models and cumulative probability models); also see Samejima (1979).

As an example of using Equation 9.4 to obtain transition points, we use the item shown in Figure 9.3. With $\alpha' = (-0.75, -0.25, 1.0)$ and $\gamma' = (-1.5, -0.25, 1.75)$, the intersection of categories 1 and 2 ($k^* = 1$ and $k = 2$) occurs at

$$\delta_{1,2} = \frac{-1.5 - (-0.25)}{-0.25 - (-0.75)} = -2.5$$

and categories 2 and 3 ($k^* = 2$ and $k = 3$) intersect at

$$\delta_{2,3} = \frac{-0.25 - 1.75}{1.0 - (-0.25)} = -1.6$$

These values match the transition points among the three ORFs shown in Figure 9.3.

The second formulation of a location parameter establishes one location parameter for each response category. Using the relationship between the intercept and location parameters from Chapter 2 (Equation 2.5, $\delta = -\gamma/\alpha$), Baker (1992) reparameterized the model in Equation 9.2 to be

$$p_{jk} = \frac{\exp(\alpha_{jk}(\theta - \delta_{jk}^\circ))}{\sum_{h=1}^{m_j} \exp(\alpha_{jh}(\theta - \delta_{jh}^\circ))} \quad (9.5)$$

where $\delta_{jk}^\circ = -\gamma_{jk}/\alpha_{jk}$. This reparameterization presents the NR model in a format similar to that seen with the dichotomous IRT models.

Applying the δ_{jk}° formulation to the example item from Figure 9.3 and dropping the item subscript, we have

$$\delta_1^\circ = \frac{-(-1.5)}{-0.75} = -2.0$$

$$\delta_2^\circ = \frac{-(-0.25)}{-0.25} = -1.0$$

and

$$\delta_3^\circ = \frac{-(1.75)}{1.0} = -1.75$$

Figure 9.3 shows that these δ_{jk}° s do not correspond to the inflection points of the ORFs, nor do they correspond to points of ORF intersection. Only with a two-category item can one interpret the δ_{jk}° s as the point on the θ scale at which an individual has a probability of 0.50 of simultaneously responding in two categories (i.e., δ_1° is the transition point between category 0 and category 1).

Thissen, Cai, and Bock (2010) reparameterized the NR to have an overall discrimination parameter, α_j , in addition to a set of α_{jk} s. In this fashion, the single α_j readily extends to be the discrimination parameter in the GPC and 2PL models. In Thissen et al.'s formulation, the $z_{jk}(\theta)$ in Equation 9.3 is

$$z_{jk}(\theta) = \gamma_{j,k+1} + \alpha_j \alpha_{j,k+1}^s \theta, \quad (9.6)$$

where α_j is item j 's overall discrimination parameter, $\gamma_{j,k+1}$ is the intercept parameter for item j , and $\alpha_{j,k+1}^s$ is the "scoring function" for response k on item j . The scoring functions for the responses are a set of $m_j - 2$ contrasts among the α_{jk}^s s (Thissen et al., 2010). To identify the model, the first α_{jk}^s and γ_{jk} are each set to 0 (i.e., $\alpha_{j1}^s = \gamma_{j1}^s = 0$) with the last α_{jk}^s equal to $m_j - 1$ (i.e., $\alpha_{jm}^s = m_j - 1$). Therefore, there are $m_j - 2$ contrasts among the α_{jm}^s s (i.e., $\underline{\alpha}' = (\alpha_{j2}^s, \dots, \alpha_{j,m-1}^s)$ and $m_j - 1$ γ_{jk} s (i.e., $\underline{\gamma}' = (\gamma_{j2}, \dots, \gamma_{jm})$). As is the case with the NR model, these constraints are implemented by using a transformation matrix, \underline{T} . However, Thissen et al.'s parameterization uses a Fourier \underline{T} in lieu of the deviation contrast approach. See Thissen et al. (2010) for more information on the extension of this model to a multidimensional context.

Information for the NR Model

As is true with the previous polytomous models, we can determine the amount of information for estimating a person's location provided by a particular item response category. For the NR model the option information function (Bock, 1972) is

$$I_{jk}(\theta) = \underline{a} \underline{W} \underline{a}' p_{jk} \quad (9.7)$$

where

$$\underline{W} = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_{m_j} \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_{m_j} \\ \vdots & \vdots & \vdots & \vdots \\ -p_{m_j}p_1 & -p_{m_j}p_2 & \dots & p_{m_j}(1-p_{m_j}) \end{bmatrix}.$$

The sum of these option information functions provides the item information function

$$I_j(\theta) = \sum_{k=1}^{m_j} \underline{a} \mathbf{W} \underline{a}' p_{jk} = \underline{a} \mathbf{W} \underline{a}'. \quad (9.8)$$

As is true for every model discussed in previous chapters, the total information provided by an instrument, $I(\theta)$, is the sum of the individual item information functions, $I_j(\theta)$.

In general, the distribution of information is affected by the distance between the item's γ_{jk} s, whether the γ_{jk} s are ordered in terms of magnitude, and the number of item alternatives (de Ayala, 1992). As is the case for the ordinal models, the information provided by the individual response categories does not have to be equal across response categories. Moreover, it is possible to obtain bimodal item information functions. The relationship between dichotomous and ordered polytomous models' information functions that we saw with the GR model also appears with the NR model. For example, Thissen's (1976) application of the NR model to the Ravens Progressive Matrices showed that the NR model provided more information than did a dichotomous model, particularly for individuals at the lower end of the scale.

Metric Transformation, NR Model

The principles outlined in previous chapters for metric conversion apply to the NR model. The intercept parameters (or their estimates) are transformed by $\gamma_{jk}^* = \gamma_{jk} - \frac{\alpha_{jk}(\kappa)}{\zeta}$ and the category slope parameters (or their estimates) by $\alpha_{jk}^* = \frac{\alpha_{jk}}{\zeta}$. Person location parameters (or their estimates) are transformed by $\theta^* = \zeta(\theta) + \kappa$. Because of the unordered nature of the response categories, it is not possible to convert the person location (estimates) to expected trait scores on the instrument without dichotomizing the responses.

Conceptual Development of the Multiple-Choice Model

With the NR model it is assumed that examinees purposefully select an item's response option. Moreover, as shown in Figure 9.3, as an examinee's location decreases, one particular response category will become increasingly more attractive to the point where it is the most likely response. For example, in Figure 9.3 this is reflected by category 1's monotonically decreasing ORF. However, in the context of proficiency assessment and a multiple-choice item format, conventional wisdom states that some individuals with very low proficiencies will randomly guess on some items. (This is the rationale that led to the development of the 3PL model.) However, according to the NR model, individuals with very low proficiencies will pick a particular option rather than randomly guess at the item's options. As such, the NR model does not address the possibility of examinee guessing behavior.

If one believes that individuals with very low θ s will randomly select an item's

option, then the observed nominal response data reflect a mixture of individuals who purposefully select the option and those who randomly select the option, with probability $1/m_j$ (i.e., the reciprocal of the item's number of options). Samejima (1979) proposed a solution to the issue of guessing in polytomous data by suggesting that these individuals who guess randomly on an item select the item's options with equal probability $1/m_j$. This idea is incorporated into extensions of both the GR and NR models (normal ogive and logistic versions) by creating a “no recognition” category (labeled 0) to reflect the individuals who do not have the proficiency to even recognize the plausibility of an item's distractor and therefore guess at random on the item. This “no recognition” category has an ORF that is strictly decreasing with respect to θ and is asymptotic with 1 and 0. However, because the individuals who belong to this category are assumed to randomly guess on the m_j options, the “no recognition” ORF “disappears” and the other options are affected by this random guessing. As such, in Samejima's approach the “no recognition” category (category 0) might be considered to be a latent response category. In the context of nominal polytomous response data, this random guessing is modeled by modifying the NR model. Samejima (1979) refers to this modified model as the *Bock–Samejima model for multiple-choice items* (BS; also known as Type I Model C).⁴ The BS model specifies the probability of an individual located at θ responding in category k of item j , given the item's $\underline{\alpha}$, $\underline{\gamma}$, and m_j as

$$p_j(x=k|\theta, \underline{\alpha}, \underline{\gamma}, m_j) = \frac{\exp(\gamma_{jk} + \alpha_{jk}\theta) + \exp(\gamma_{j0} + \alpha_{j0}\theta)(1/m_j)}{\sum_{h=0}^{m_j} \exp(\gamma_{jh} + \alpha_{jh}\theta)} \quad (9.9)$$

where $\alpha_{j0} < \alpha_{j1} < \dots < \alpha_{jm_j}$. The fixed proportion, $1/m_j$, is incorporated into each of the observed categories by the second term in the numerator of Equation 9.9, and if $\alpha_{j0} = \gamma_{j0} = 0$, then the model simplifies to the NR model (Thissen, Steinberg, & Fitzpatrick, 1989).

A second approach to modeling nominal polytomous data in the presence of guessing is presented by Thissen and Steinberg (1984). As is the case with the Bock–Samejima model, they conceptualized a latent response category for the “no recognition” individuals; they refer to these individuals as “don't know” individuals. However, in contrast to Samejima, they felt that these “don't know” individuals would not guess with the equal probability of $1/m_j$ and they presented data that supported their contention. Therefore, they modified Equation 9.9 to allow the proportion of “don't know” individuals to vary across an item's options. Their model is called the *multiple-response model* or the *multiple-choice (MC) model*. According to the MC model, the probability of a person located at θ responding in the k th category of item j is given by

$$p_j(x=k|\theta, \underline{\alpha}, \underline{\gamma}, \phi) = \frac{\exp(\gamma_{jk} + \alpha_{jk}\theta) + \phi_{jk}(\exp(\gamma_{j0} + \alpha_{j0}\theta))}{\sum_{h=0}^{m_j} \exp(\gamma_{jh} + \alpha_{jh}\theta)} \quad (9.10)$$

where m_j , θ , α_{jk} and γ_{jk} are defined above, $k = \{1, \dots, m_j\}$, and the model's new parameter,

ϕ_{jk} , is associated with the latent “don’t know” response category and is labeled 0 (i.e., α_{j0}, γ_{j0}). This unobserved response category is sometimes also referred to as category 0 (Thissen & Steinberg, 1997). For convenience we use p_{jk} for $p_j(x = k|\theta, \alpha, \gamma, \phi)$ in the following.

To summarize, in certain situations (e.g., proficiency assessment), the respondents to an item are assumed to consist of a mixture of two classes of individuals: those who choose specific item options and those who, because they “don’t know,” randomly choose among the response categories. Therefore, conceptually one is faced with the task of “unmixing” the respondents for each response category. The ϕ_{jk} parameter represents the proportion of respondents who “don’t know” and select category k on item j ; ϕ_{jk} is a function of the estimated parameters (Thissen & Steinberg, 1984; Thissen et al., 1989a). As is done with the α_{jk} s and the γ_{jk} s, we can collect the $m_j \phi_{jk}$ s into a vector $\phi = (\phi_{j1}, \phi_{j2}, \dots, \phi_{jm})$.

The model presented in Equation 9.10 is in terms of the transformed α_{jk} s, γ_{jk} s, and ϕ_{jk} s. That is, similar to the case with the NR model, the α_{jk} s, γ_{jk} s, and ϕ_{jk} s in Equation 9.10 are the result of transformations to address the indeterminacy issue and identify the model. These transformations of the estimated (unconstrained) parameters impose the constraints

$$\sum_{k=0}^{m_j} \alpha_{jk} = 0, \quad \sum_{k=0}^{m_j} \gamma_{jk} = 0, \quad \text{and} \quad \sum_{k=0}^{m_j} \phi_{jk} = 0.$$

As is the case with the NR model, the first two constraints are imposed by applying the transformation matrix, \mathbf{T}_ϕ to the unconstrained slope, α_{jk}^u , and intercept, γ_{jk}^u , parameters (i.e., $\alpha = \alpha^u \mathbf{T}$ and $\gamma = \gamma^u \mathbf{T}$). The last constraint is imposed using a transformation matrix. This matrix, \mathbf{T}_ϕ , is applied to the unconstrained “don’t know” parameters, ϕ_{jk} s (i.e., $\phi = \phi^u \mathbf{T}_\phi$); \mathbf{T} and \mathbf{T}_ϕ have different dimensions. As a result, the number of unconstrained (free) parameters per item is $3m_j - 1$ (i.e., $m_j \alpha_{jk}$ s, $m_j \gamma_{jk}$ s, and $(m_j - 1) \phi_{jk}$ s). In addition, for each item j there are $(m_j + 1)$ constrained α_{jk} s ($m_j + 1$) constrained γ_{jk} s, and m_j constrained ϕ_{jk} s. If ϕ^u is fixed as a null vector (i.e., \emptyset), then each of the ϕ_{jk} s equals $1/m_j$ and the MC model simplifies to the BS model (Thissen & Steinberg, 1984). (A calibration example, MULTILOG_MCMcalibrationEx.pdf, may be found on the author’s website.)

How Large a Calibration Sample?

Because of the interaction between the α_{jk} s and the γ_{jk} s, as well as the effect of the distribution of respondents across categories, calibration sample size recommendations should be seen simply as rough guidelines. One guideline suggests a minimum ratio of 10:1 respondents to the total number of item parameters when the respondents are normally distributed, but a larger ratio if the respondents are not normally distributed (de Ayala & Sava-Bolestá, 1999).

DeMars (2003) investigated parameter recovery using two sample sizes (600 and 2400), two levels of the number of parameters per item (6 and 12; that is, three- and

six-response category items), two levels of the total number of item parameters across all items, and three different respondent population distributions (normal, skewed, uniform). Her results showed that, on average, the estimation of γ_{jk} s was better than that of the α_{jk} s, decreasing the number of item parameters per item from 12 to 6 led to more accurate α_{jk} and γ_{jk} estimates, and the 2400 sample size led to smaller root mean squared error (RMSE) than did a sample size of 600 for corresponding conditions. She suggested that one should focus on the ratio of sample size to number of categories, although she did not give a specific ratio because of the complex nature of the relationship of a sample's distribution to locations, category discrimination, and the degree of acceptable error. The DeMars (2003) study replicated the de Ayala and Sava-Bolesta (1999) finding that normally distributed respondents led to, overall, the best results. In this regard, DeMars showed that it is possible to obtain acceptable estimation accuracy ($\text{RMSE} < 0.10$) for both the α_{jk} s and the γ_{jk} s with 600 respondents, provided one has a normal (or a uniform) distribution of respondents and three-category items. However, DeMars's best results (i.e., $\text{RMSE} < 0.10$) for both α_{jk} s and γ_{jk} s were obtained with sample size ratios of 10:1 and 20:1 and either normally or uniformly distributed respondents. The de Ayala and Sava-Bolesta and DeMars studies show that the match between the prior distribution used in estimation and the respondents' distribution is important for accurately estimating the item parameters. A similar finding was obtained by Wollack, Bolt, Cohen, and Lee (2002). For those cases where respondents are not normally distributed, modifying the prior distribution used in estimation may improve the accuracy of the estimated item parameters.

As we have done in previous chapters, we provide very rough sample size guidelines. Assuming MMLE, a symmetric θ distribution, and that the respondents distribute themselves across the response categories in reasonable numbers, we suggest the minimum sample size be 600. However, it may be anticipated that there is a sample size, say 1500 or so, at which one reaches, practically speaking, a point of diminishing returns in terms of improvement in estimation accuracy. (The value of 1500 should not be interpreted as an upper bound.) For instance, we conducted a small illustrative simulation in which three sample sizes were generated for a four-item instrument; $m_j = 4$. With four 4-choice items there are 256 possible response patterns, and for each sample size all the patterns were observed. Increasing the sample size from 973 cases to 1799 resulted in the corresponding item parameter estimates changing by an amount from 0 up to 0.07. In general, most of the changes between the two calibrations were on the order of 0.01 to 0.03. However, "doubling" the sample size from 1799 to 3599 yielded no change in the item parameter estimates, thereby providing some support for the diminishing returns assertion.

If one adopts a sample size ratio for sample size determination (e.g., 10 persons for every parameter estimated), then it is probably more useful closer to the lower bound than to the value of 1500 (i.e., when the sample size is large, then the sample size ratio becomes less important). These suggestions are tempered by the purpose of the administration (e.g., survey, establishing norms, equating, item pool development/maintenance), the application characteristics (e.g., distribution and range of item parameter estimates, instrument length, latent distribution shape), ancillary technique sample size

requirements, and the amount of missing data. As previously mentioned, sample size guidelines should not be interpreted as hard-and-fast rules.

Example: Application of the NR Model to a General Science Test, MMLE, mirt

The data for our example come from the Third International Mathematics and Science Study (TIMSS; Gonzalez et al., 1998) database. Our responses to five science items come from the 1995 Canadian examinees subset ($N = 14,611$). All of our items use a four-option multiple-choice format, with two items covering earth science and one each covering life science, physical science, and environmental issues/nature of science. The responses to each item are coded 1 through 4. It should be noted that the toolbox of model–data fit methods summarized in Chapter 6 is still relevant and would be used in practice. We assume conditional independence and a unidimensional latent space; Johnson and Bolt (2010) present a multidimensional extension of the NR model.

Several programs are available for NR model estimation (e.g., flexMIRT, mirt, TAM).⁵ We use mirt to calibrate our data. Our R session is shown in Table 9.1. We verify that our data were correctly read by using the head and tail functions. After removing the id variable, we use the describe function from the Hmisc package (describe(condomsdata)). Our results show that the correct number of response categories for each item, that N is correct, that our correct option always has the largest frequency, and that we do not have any categories with very small or zero frequencies.

Chalmers (2019, p. 106) recommends that we “choose high and low anchors that cause the estimated parameters to fall between 0 and $K-1$ either by theoretical means or by re-estimating the model with better values following convergence.” The “nominal model can become numerical[ly] unstable”; Chalmers’s K is m_j . Our solution to this problem is a simple and nonmodel base. (By using a nonmodel base strategy, we avoid model/item misfit adversely affecting the starting values.) To provide starting values, we use the attractiveness of the items’ options. We indicate option attractiveness by assigning the value $m_j - 1$ to the most attractive option, 0 to the least attractive option, and the rank order to the remaining options (i.e., we use a 0-based system). As an example, consider item 1 and its frequency distribution. In terms of option attractiveness, we have option 2 (the correct option), followed by options 3, 4, and 1. Therefore, we reflect option 2 with the value 3 ($m_j - 1 = 4 - 1 = 3$), option 1 with 0, and options 3 and 4 with 2 and 1, respectively. mirt uses Thissen et al.’s (2010) scoring function parameterization of the NR model. Therefore, in terms of scoring functions and mirt’s notation, we have m_j parameter labels (e.g., ak) whose ordinal positions are 0 to $m_j - 1$. mirt’s syntax is the keyword START, the assignment operator “=”, followed by a parenthesized term of the item number, parameter label, and the value to be assigned. For instance, for item 1 we want ak0 (i.e., option 1) set to 0 (START = (1, ak0, 0)), ak1 (i.e., option 2) set to 3 (START = (1, ak1, 3)), ak2 (option 3) set to 2 (START = (1, ak2, 2)), and ak3 (option 4) set to 1 (START = (1, ak3, 1)). In a similar way the remaining four items are defined. All of this information is assigned to the freqbasis object;

TABLE 9.1. mirt Session for the NR Model Calibration of the General Science Data

```

> library(mirt); library(car); library(Hmisc)
> packageVersion("car")
[1] '3.0.3'
> packageVersion("Hmisc")
[1] '4.3.0'

> GenSciencedata=read.table("GenScience.dat",col.names=c("id",paste0("i",1:5)))

> head(GenSciencedata,5)
  id i1 i2 i3 i4 i5
1 1  2  1  1  3  4
2 2  2  2  1  4  4
3 3  2  1  1  4  4
4 4  2  1  1  4  4
5 5  3  3  1  4  4

> tail(GenSciencedata,5)
  id i1 i2 i3 i4 i5
14607 14607 2 1 1 4 4
14608 14608 2 1 1 4 4
14609 14609 2 1 1 4 4
14610 14610 2 3 1 4 4
14611 14611 2 2 1 4 4

> GenSciencedata=within(GenSciencedata,rm(id)) # drop id variable

> Hmisc::describe(GenSciencedata)
  GenSciencedata

  5 Variables      14611 Observations
-----
i1
  n missing distinct      Info      Mean      Gmd
  14611      0        4   0.345    2.132   0.2952

  Value      1      2      3      4
  Frequency  219 12686 1258  448
  Proportion 0.015 0.868 0.086 0.031
-----
i2
  n missing distinct      Info      Mean      Gmd
  14611      0        4   0.511    1.427   0.7139

  Value      1      2      3      4
  Frequency 11500 1156  789  1166
  Proportion 0.787 0.079 0.054 0.080
-----
i3
  n missing distinct      Info      Mean      Gmd
  14611      0        4   0.601    1.656   1.016

  Value      1      2      3      4
  Frequency 10706 737   658  2510
  Proportion 0.733 0.050 0.045 0.172
-----
i4
  n missing distinct      Info      Mean      Gmd
  14611      0        4   0.728    3.318   0.996

```

(continued)

TABLE 9.1. (continued)

Value	1	2	3	4		
Frequency	1741	1271	2195	9404		
Proportion	0.119	0.087	0.150	0.644		
<hr/>						
i5	n	missing	distinct	Info	Mean	Gmd
	14611	0	4	0.367	3.641	0.6297
<hr/>						
Value	1	2	3	4		
Frequency	1430	311	328	12542		
Proportion	0.098	0.021	0.022	0.858		
<hr/>						
> freqbasis = 'F=1-5						
+ START = (1, ak0, 0)						
+ START = (1, ak1, 3)						
+ START = (1, ak2, 2)						
+ START = (1, ak3, 1)						
+ START = (2, ak0, 3)						
+ START = (2, ak1, 1)						
+ START = (2, ak2, 0)						
+ START = (2, ak3, 2)						
+ START = (3, ak0, 3)						
+ START = (3, ak1, 1)						
+ START = (3, ak2, 0)						
+ START = (3, ak3, 2)						
+ START = (4, ak0, 1)						
+ START = (4, ak1, 0)						
+ START = (4, ak2, 2)						
+ START = (4, ak3, 3)						
+ START = (5, ak0, 2)						
+ START = (5, ak1, 0)						
+ START = (5, ak2, 1)						
+ START = (5, ak3, 3)'						
<hr/>						
> print((nrm=mirt(data=GenSciencedata,model=freqbasis,itemtype="nominal",SE=T)))						
Iteration: 40, Log-Lik: -51781.058, Max-Change: 0.00007						
<hr/>						
Calculating information matrix...						
<hr/>						
Call:						
mirt(data = GenSciencedata, model = freqbasis, itemtype = "nominal",						
SE = T)						
<hr/>						
Full-information item factor analysis with 1 factor(s).						
Converged within 1e-04 tolerance after 40 EM iterations.						
mirt version: 1.30						
M-step optimizer: BFGS						
EM acceleration: Ramsay						
Number of rectangular quadrature: 61						
Latent density type: Gaussian						
Information matrix estimated with method: Oakes						
Condition number of information matrix = 887.6916						
Second-order test: model is a possible local maximum						
<hr/>						
Log-likelihood = -51781.06						
Estimated parameters: 30						
AIC = 103622.1; AICC = 103622.2						
BIC = 103849.8; SABIC = 103754.5						
G2 (993) = 1173.57, p = 1e-04						
RMSEA = 0.004, CFI = NaN, TLI = NaN						

(continued)

TABLE 9.1. (continued)

```

> # iteration history
> print((as.data.frame((extract.mirt(nrm, 'LLhistory')))), digits=10)
  (extract.mirt(nrm, "LLhistory"))
  1           -71258.62342
  2           -59738.89291
  3           -54603.46292
  4           -53030.14650
  5           -52061.90715
  :
  38          -51781.05834
  39          -51781.05812
  40          -51781.05812

> itemfit(nrm, empirical.table=2, group.bins=8)
$`theta = -1.2901`
  Observed Expected z.Residual
cat_1      340 851.4740 -17.528222
cat_2      527 367.2915  8.333406
cat_3      567 276.7798 17.444580
cat_4      393 331.4547  3.380513

$`theta = -0.6162`
  Observed Expected z.Residual
cat_1      804 1298.6078 -13.725304
cat_2      404 199.0742 14.524097
cat_3      222 118.1967  9.547911
cat_4      396 210.1212 12.823152

$`theta = -0.2209`
  Observed Expected z.Residual
cat_1     1224 1492.61648 -6.952781
cat_2      225 124.70651  8.981072
cat_3       0  64.37823 -8.023604
cat_4      377 144.29878 19.371681

$`theta = 0.0185`
  Observed Expected z.Residual
cat_1     1826 1579.49226  6.202572
cat_2       0  91.37860 -9.559215
cat_3       0  43.34255 -6.583506
cat_4       0 111.78659 -10.572918

$`theta = 0.2939`
  Observed Expected z.Residual
cat_1     1826 1654.50769  4.216094
cat_2       0  62.71400 -7.919217
cat_3       0  26.98486 -5.194695
cat_4       0  81.79345 -9.043973

$`theta = 0.5883`
  Observed Expected z.Residual
cat_1     1826 1711.08172  2.778138
cat_2       0  41.27287 -6.424396
cat_3       0  16.00245 -4.000306
cat_4       0  57.64297 -7.592296

```

(continued)

TABLE 9.1. (continued)

```
$`theta = 0.6134`  

  Observed   Expected z.Residual  

cat_1      1826 1714.97984  2.680849  

cat_2       0    39.80509 -6.309128  

cat_3       0    15.29709 -3.911150  

cat_4       0    55.91798 -7.477833  

$`theta = 0.6134`  

  Observed   Expected z.Residual  

cat_1      1828 1716.85824  2.682317  

cat_2       0    39.84869 -6.312582  

cat_3       0    15.31385 -3.913291  

cat_4       0    55.97923 -7.481927  

> itemfit(nrm,group.bins=8,empirical.plot=2,empirical.CI=0))      # produces Figure 9.4  

> itemfit(nrm,fit_stats="X2*_df")  

  item   X2_star_scaled df.X2_star_scaled RMSEA.X2_star_scaled p.X2_star_scaled  

  1   i1        227.104      197.444      0.003      0.073  

  2   i2        121.469      132.750      0.000      0.749  

  3   i3        73.001       75.306      0.000      0.554  

  4   i4        99.123       85.032      0.003      0.141  

  5   i5       137.593      142.392      0.000      0.598  

> coef(nrm,  simplify=T, IRTpars=T)  

  $items  

    a1     a2     a3     a4     c1     c2     c3     c4  

i1 -0.507  0.816 -0.472  0.164 -1.905  2.738 -0.124 -0.709  

i2  1.182 -0.353 -0.707 -0.121  2.252 -0.570 -1.309 -0.373  

i3  0.607 -0.417 -0.369  0.179  1.885 -1.112 -1.189  0.416  

i4 -0.304 -0.476  0.116  0.664 -0.454 -0.881 -0.055  1.390  

i5  0.255 -0.697 -0.773  1.216  0.648 -1.835 -1.885  3.072  

  $means  

  F  

  0  

  $cov  

  F  

  F 1  

> # ORFs item 1 (Figure 9.5)  

> plot(nrm, type = 'trace', which.items = 1,theta_lim=c(-4,4), auto.key =  

  list(points=FALSE,lines=TRUE, columns=4),par.settings = simpleTheme(lty=1:4))  

> # item 1 info (Figure 9.5)  

> plot(nrm, type = 'infotrace', which.items = 1,theta_lim=c(-4,4),scales=list  

  (tick.number=10))  

> # ORFs all items (Figure 9.6, top)  

> plot(nrm, type = 'trace', which.items = c(3,4,5,1,2),theta_lim=c(-4,4), auto.key =  

  list(points=FALSE,lines=TRUE, columns=4),par.settings = simpleTheme(lty=1:4))  

> # ORFs all items (Figure 9.6, bottom)  

> plot(nrm, type = 'infotrace', which.items = c(3,4,5,1,2),theta_lim=c(-4,4),  

  scales=list(tick.number=10))  

> # Total information (Figure 9.7)  

> plot(nrm, type = 'info',theta_lim=c(-4,4),scales=list(tick.number=10))
```

because we are assuming a unidimensional space, all five items are loading on a single factor ($F = 1-5$).

To perform our calibration, we specify the nominal model as our itemtype and pass freqbasis to the mirt function (`nrm = mirt(mathdata, model = freqbasis, itemtype = 'nominal')`). Our calibration requires 40 iterations to obtain convergence. The iteration history shows the desired progression of smaller and smaller changes in the log likelihood, culminating in a $\ln L$ of -51,781.058 with $AIC = 103,622.1$ and $BIC = 103,849.8$.

As an example of item fit analysis, we use `itemfit` with item 2 (`itemfit(nrm, empirical.table = 2, group.bins = 8)`). These results are interpreted similar to those discussed above (e.g., Chapters 4 and 5). For instance, for the first fractile (“examinees located at -1.2901”) we expected to see approximately 851 individuals to select the first option (the correct answer), but only 340 did so. Conversely, first fractile examinees tended to select options 2–4 more than we would predict. Stated another way, examinees in the first fractile tended to select incorrect options more than we would predict. As we examine the other fractiles (i.e., θ increases), we see an increasing number of examinees selecting the first option and a decreasing number selecting options 2 through 4. As can be seen, the residuals are “largish.” To further examine this, we invoke the `itemfit` function with the `empirical.plot` argument (Figure 9.4). For the reasons discussed above, we do not weight the lowest and highest bins as much

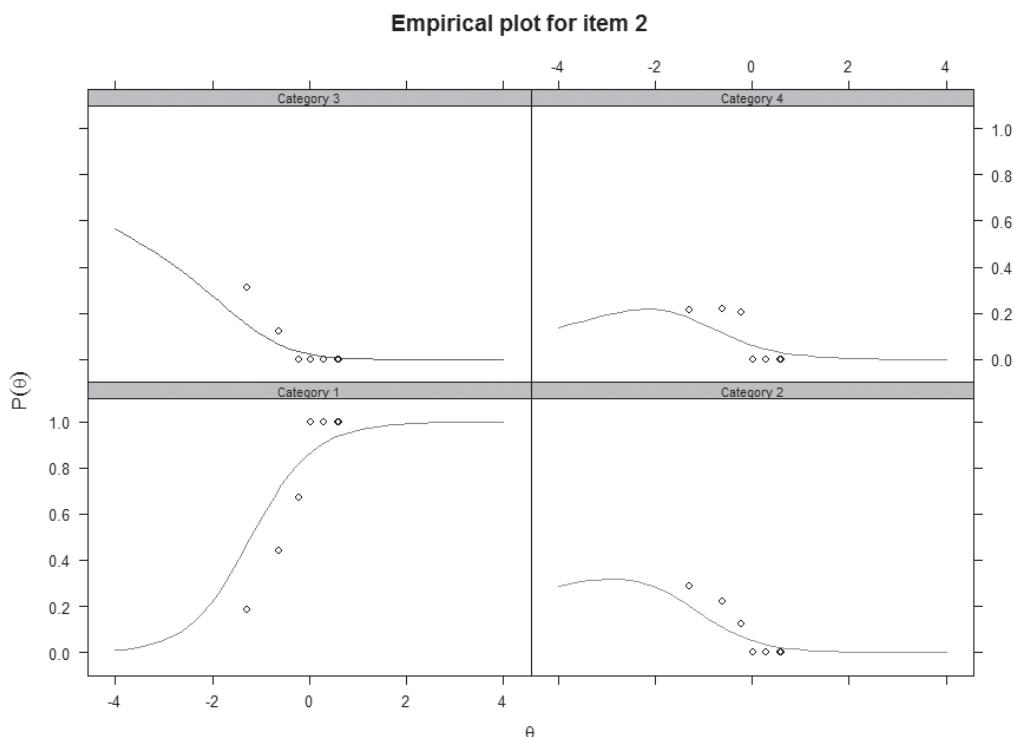


FIGURE 9.4. Predicted and observed ORFs for item 2.

as those in toward the center. Comparing the empirical and predicted ORFs shows the trace lines generally follow what is observed, although there are some bins for which the response function is a poor approximation. However, there is nothing patently egregious that would cause us not to proceed to look at our statistical fit indices. This figure also illustrates the difficulty in estimating option parameters when the examinees are not distributed throughout the θ range in sufficient numbers (e.g., having to extrapolate in both directions with little to no information).

We use Stone's (2000) rescaled χ^2 * as our item statistical fit index (`itemfit(nrm, fit_stats = "X2* _ df")`). As mentioned in Chapter 5, χ^2 * performs well in terms of power while maintaining a nominal Type I error rate in line with the significance level.⁶ For our first item we have a rescaled χ^2 * (χ_s^{2*}) of 227.104 (`x2_star_scaled`) on 197.444 `dfs` (`df.X2_star_scaled`) with a $p(\chi_s^{2*} | H_0 \text{ true}) = 0.073$ (`p.X2_star_scaled`) and a $\text{RMSEA} = 0.003$ (`RMSEA.X2_star_scaled`). For all five items we fail to reject the null hypothesis that the data are consistent with the NR model at ($\alpha = 0.05$). Following MacCallum, Browne, and Sugawara (1996), our items' RMSEAs reflect "close fit." Thus, we proceed with our analysis.

We obtain our item parameter estimates using the `coef` function (e.g., `coef(nrm, simplify = T, IRTpars = T)`); these estimates reflect Bock's parameterization.⁷ The constrained category discrimination parameter estimates are found in the `a1, a2, ..., a4` columns. That is, for item 1 we have $\hat{\alpha}_{11} = -0.507$, $\hat{\alpha}_{12} = 0.816$, $\hat{\alpha}_{13} = -0.472$, and $\hat{\alpha}_{14} = 0.164$ or $\hat{\alpha}' = (-0.507, 0.816, -0.472, 0.164)$. Similarly, the item 1's constrained intercept estimates are listed in the `c1, c2, ..., c4` columns $\hat{\gamma}_{11} = -1.905$, $\hat{\gamma}_{12} = 2.738$, $\hat{\gamma}_{13} = -0.124$, and $\hat{\gamma}_{14} = -0.709$ or $\hat{\gamma}' = (-1.905, 2.738, -0.124, -0.709)$.⁸ For each of the items, the largest $\hat{\alpha}_{jk}$ is associated with the correct response. As would be expected, the sum of the constrained $\hat{\alpha}_{jk}$ s for an item is 0.0, as is the case for an item's $\hat{\gamma}_{jk}$ s. Looking at item 1, one sees that category 2 does the best of the item's response categories in discriminating among the individuals ($\hat{\alpha}_{12} = 0.816$), and it is also the most attractive ($\hat{\gamma}_{12} = 2.738$). Because the γ s are associated with category frequencies, the category with the largest frequency has the largest positive γ , and the category with smallest frequency is associated with the largest negative γ . (If a category has its α_{jk} equal to zero, then its corresponding γ is poorly defined and has a tendency to drift. In this situation the calibration may not converge, but according to Thissen [1982] there is little loss of fit by allowing the calibration to stop because of reaching the maximum number of iterations.)

In some situations, one or more of an item's response categories may not be attractive and may never be chosen. These are sometimes referred to as *null* categories. In these cases, one does not have data to estimate the category's parameters. In short, the item is functioning with fewer categories than are specified for the calibration. This situation would reveal itself by the null category's corresponding observed proportion and frequency being zero, as well as by the absence of an ORF for the null category in the item's ORF plot. If a null category occurs, then one should ignore the null category's parameter estimates and recalibrate the item set specifying the appropriate number of observed categories for each item. For instance, assume that item 4's fourth option is not selected by any individuals. Therefore, this item is functioning as a three-category item.

Item 1's ORFs (left panel) and its information (right panel) are shown in Figure 9.5. The correct option is shown by the monotonic increasing response function (dash line). Generally, examinees located above approximately -2.2 tend to select the correct answer, whereas below this point category 3 is most attractive, although some of these examinees tend to select categories 1 or 4. The intersection between categories 2 and 3 is given by Equation 9.4:

$$\hat{\delta}_{2,3} = \frac{2.738 - (-0.124)}{-0.472 - 0.816} = -2.222$$

The item's information function (Figure 9.5, right panel) shows that this item is most useful for estimating persons located around -2 .

Figure 9.6 shows the ORFs and item information functions for all items. For each, the correct response is reflected in the monotonically increasing response function. Moreover, one sees that the response functions tend to intersect below $\theta = 0$, indicating that examinees located below 0 are the ones primarily selecting the incorrect alternatives. In terms of estimating person locations, we see that the corresponding item information functions show their maxima below $\theta = 0$. This shows that the information gleaned from the incorrect options tends to come from individuals located in the lower half of the continuum. Not surprisingly, the total information function is skewed right with its maximum located around -2 (Figure 9.7).

Examining the ORFs (Figure 9.6), one sees that item 1's categories 1 and 4 are not as attractive to examinees as categories 2 (correct option) and 3. Therefore, for pedagogical reasons we ignore any beneficial diagnostic information differences between the options and we decide to re-code item 1 to be a dichotomy. Table 9.2 shows the R session. Using the `recode` function (from the `car` package) we re-code all responses of 3 and 4 to be 1. We double check the re-coding by obtaining the recoded item 1's frequency distribution (`with(. . . , table(i1))`) and compare it with that obtained above (Table 9.1). In contrast to the first calibration requiring 40 iterations to achieve convergence, our second calibration converged in 30 iterations with a decrease in $\ln L$ and the infor-

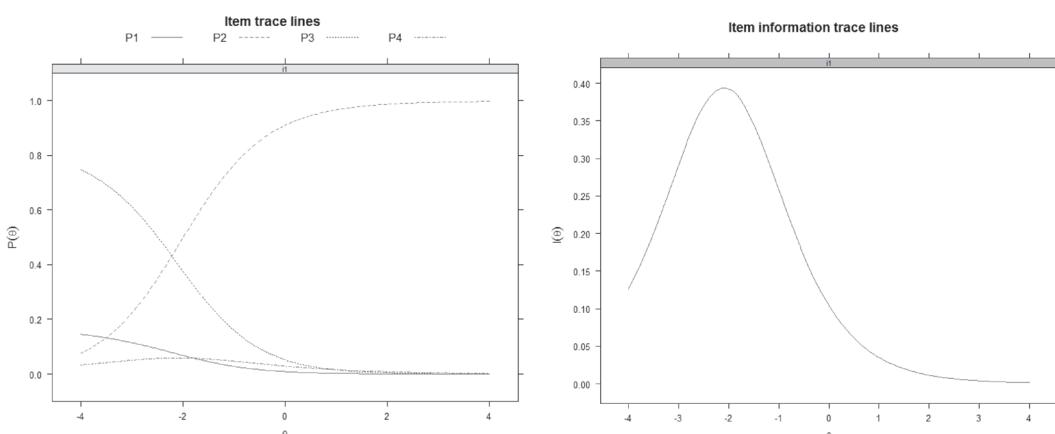


FIGURE 9.5. ORFs (left) and item information function (right) for item 1.

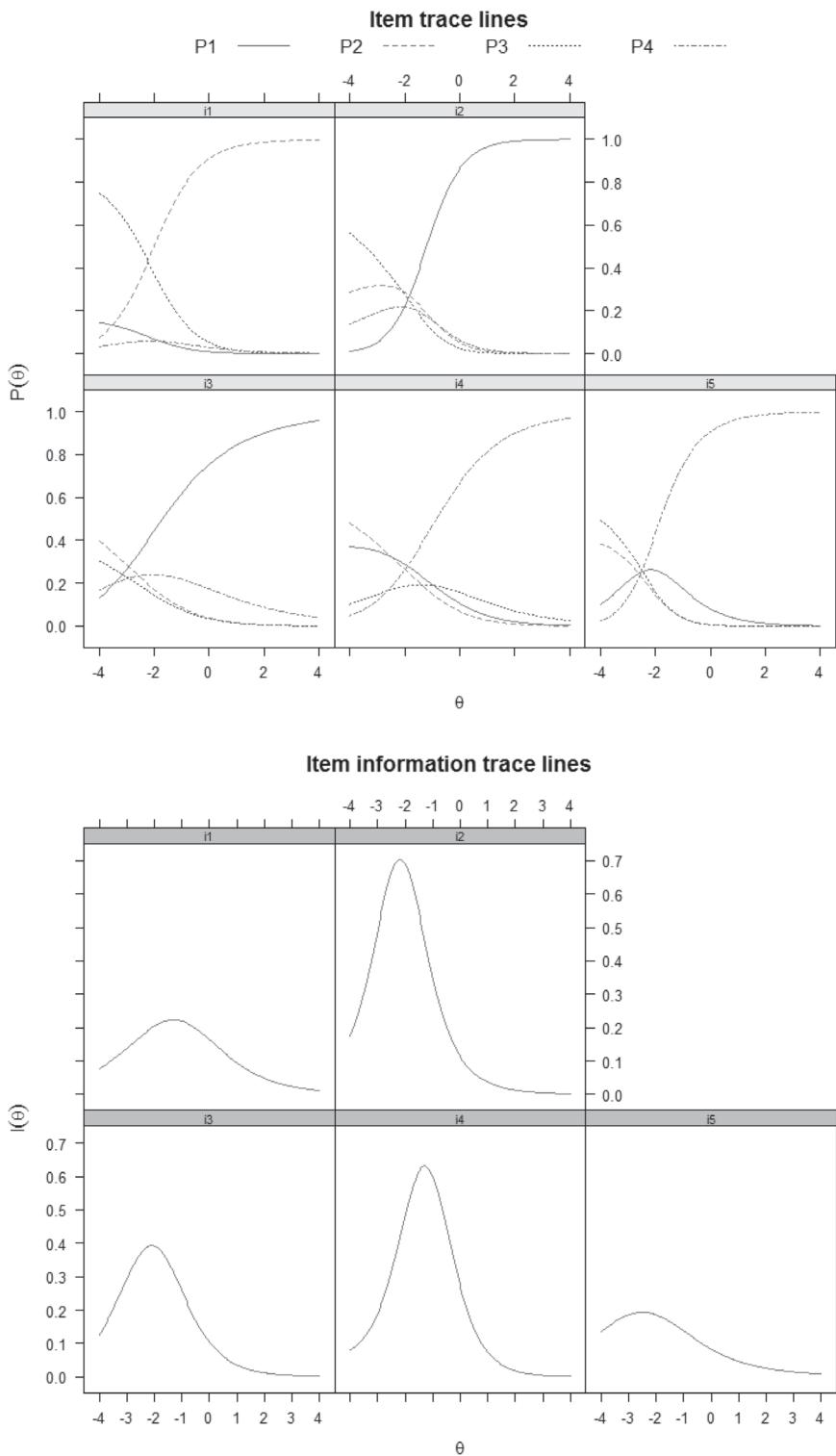


FIGURE 9.6. ORFs (top) and item information (bottom) for all five items.

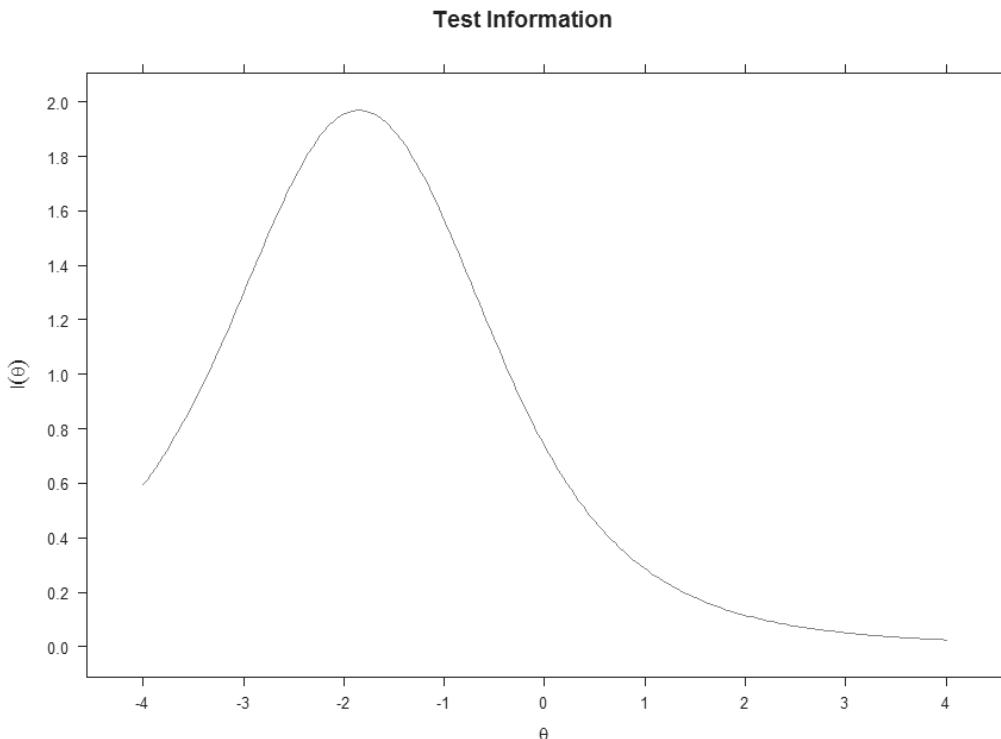


FIGURE 9.7. Total information.

mation criteria ($\ln L = -50,142.2$, AIC = 100,336.4, BIC = 100,533.7). The lack of item-level statistical misfit reflected in the above nonsignificant $\chi^2_s^{*}$ s is realized with our dichotomization of item 1 (i.e., dichotomizing does not adversely impact our $\chi^2_s^{*}$ s). The ORFs for all items are shown in Figure 9.8. In general, the pattern seen for items 2 to 5 above (Figure 9.6) is evident with our new plot. Of course, the only change is with item 1 with its two response functions represented by its IRF. The general pattern among the $\hat{\alpha}_{jk}$ s and $\hat{\gamma}_{jk}$ s seen above is reflected in our new estimates; Figure 9.8 shows items 2–5's ORFs and item 1's IRF. The only substantive difference in our estimates from above occurs with item 1. As would be expected, with only incorrect and correct responses we only have two $\hat{\alpha}_{jk}$ s and two $\hat{\gamma}_{jk}$ s. By treating item 1 as dichotomous, we are fitting the 2PL model to its data. (Technically, this is a mixed-model calibration.⁹) Item 1's parameter estimates are $\hat{\alpha}_{11} = -0.569$, $\hat{\alpha}_{12} = 0.569$, $\hat{\gamma}_{11} = -1.152$, and $\hat{\gamma}_{12} = 1.152$. We can transform our category discrimination and intercept parameter estimates to the 2PL model's α_j and γ_j by separately applying Equation 9.4's denominator and numerator, respectively.¹⁰ That is,

$$\alpha_j = \alpha_{j2} - \alpha_{j1} \text{ and} \quad (9.12)$$

$$\gamma_j = \gamma_{j1} - \gamma_{j2}. \quad (9.13)$$

TABLE 9.2. mirt Session for the NR Model Calibration of the General Science Data Treating Item 1 as a Dichotomy

```

:
> # This is a continuation of the session from Table 9.1

> # collapsing item 1's cats 1, 3, & 4 together
> #      change 1, 3, 4 to 0s (actually 1s) & # 2 to be 1(actually 2s)

> GenSciencedatai1m2=GenSciencedata
> GenSciencedatai1m2$i1=recode(GenSciencedatai1m2$i1,"3=1;4=1")

> with(GenSciencedatai1m2,table(i1))
  i1
    1      2
  1925 12686

># no need to do item 1 because already in order & is binary
> freqbasis = 'F=1-5
+ START = (2, ak0, 3)
+ START = (2, ak1, 1)
+ START = (2, ak2, 0)
+ START = (2, ak3, 2)
+ START = (3, ak0, 3)
+ START = (3, ak1, 1)
+ START = (3, ak2, 0)
+ START = (3, ak3, 2)
+ START = (4, ak0, 1)
+ START = (4, ak1, 0)
+ START = (4, ak2, 2)
+ START = (4, ak3, 3)
+ START = (5, ak0, 2)
+ START = (5, ak2, 1)
+ START = (5, ak3, 3)'

> print((nrm1m2=mirt(data=GenSciencedatai1m2,model=freqbasis,itemtype="nominal",
  SE=T)))
Iteration: 30, Log-Lik: -50142.205, Max-Change: 0.00009

  Calculating information matrix...

  Call:
  mirt(data = GenSciencedatai1m2, model = freqbasis, itemtype = "nominal",
        SE = T)

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 30 EM iterations.
  mirt version: 1.30
  M-step optimizer: BFGS
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Information matrix estimated with method: Oakes
  Condition number of information matrix = 454.2045
  Second-order test: model is a possible local maximum

  Log-likelihood = -50142.2
  Estimated parameters: 26
  AIC = 100336.4; AICC = 100336.5

```

(continued)

TABLE 9.2. (continued)

```

BIC = 100533.7; SABIC = 100451.1
G2 (485) = 633.08, p = 0
RMSEA = 0.005, CFI = NaN, TLI = NaN

> coef(nrmi1m2, simplify=T, IRTpars=T)
$items
    a1     a2     c1     c2     a3     a4     c3     c4
i1 -0.569  0.569 -1.152  1.152   NA     NA     NA     NA
i2  1.177 -0.361  2.249 -0.577 -0.705 -0.110 -1.306 -0.365
i3  0.609 -0.422  1.886 -1.115 -0.370  0.184 -1.189  0.418
i4 -0.298 -0.474 -0.452 -0.880  0.112  0.660 -0.057  1.389
i5  0.262 -0.699  0.659 -1.837 -0.785  1.223 -1.902  3.081

$means
F
0

$cov
F
F 1

> itemfit(nrmi1m2, fit_stats="X2*df")
      item X2_star_scaled df.X2_star_scaled RMSEA.X2_star_scaled p.X2_star_scaled
1     i1       191.897        195.094        0.000        0.551
2     i2       120.366        126.674        0.000        0.641
3     i3        99.151        99.478        0.000        0.490
4     i4        95.994        81.264        0.004        0.126
5     i5       148.454        147.394        0.001        0.460

> # ORFs all items (Figure 9.8)
> plot(nrmi1m2, type = 'trace', which.items = c(3,4,5,1,2), theta_lim=c(-4,4),
      auto.key = list(points=FALSE, lines=TRUE, columns=4), par.settings =
      simpleTheme(lty=1:4))

> anova(nrmi1m2, nrm)
Model 1: mirt(data = GenSciencedata, model = freqbasis, itemtype = "nominal",
SE = T)
Model 2: mirt(data = GenSciencedata1m2, model = freqbasis, itemtype = "nominal",
SE = T)

      AIC      AICC      SABIC       HQ       BIC      logLik       X2      df      p
1 103622.1 103622.2 103754.5 103697.8 103849.8 -51781.06      NaN  NaN  NaN
2 100336.4 100336.5 100451.1 100402.0 100533.7 -50142.21 3277.706 508  0

> # item info for item 1 - polytomous (Figure 9.9 - left)
> plot(nrm, type = 'infotrace', which.items = 1, theta_lim=c(-4,4),
      scales=list(tick.number=10), ylim=c(0,0.5))

> # item info for item 1 - dichotomy (Figure 9.9 - right)
> plot(nrmi1m2, type = 'infotrace', which.items = 1, theta_lim=c(-4,4),
      scales=list(tick.number=10), ylim=c(0,0.5))

```

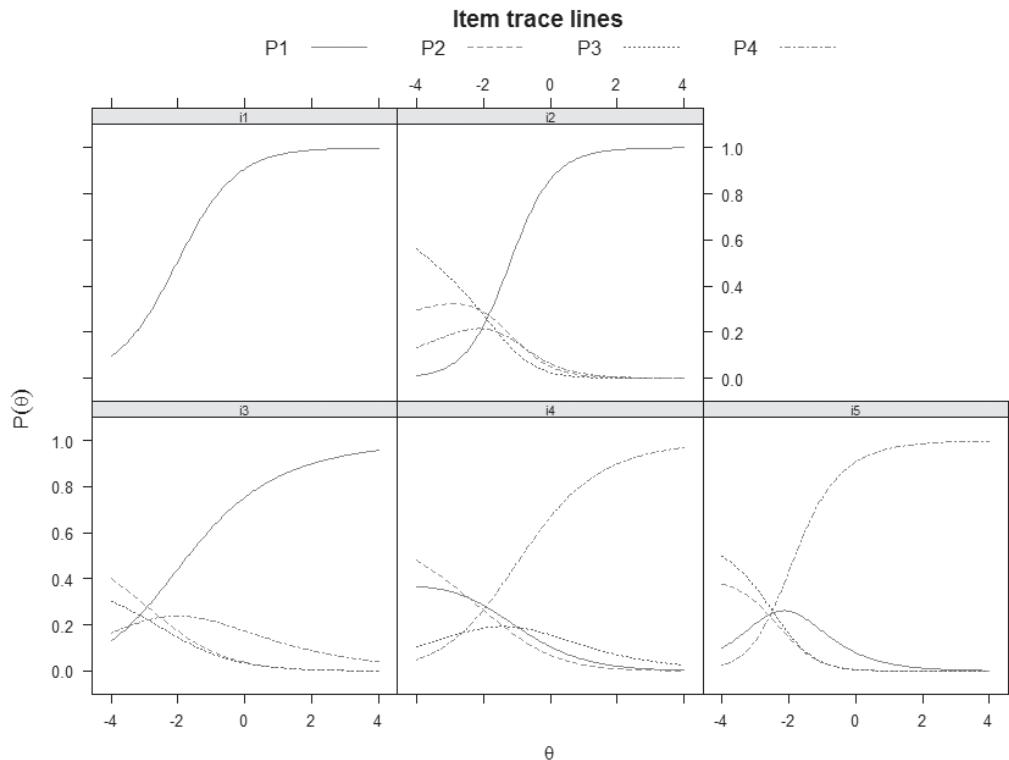


FIGURE 9.8. ORFs for all five items with Item 1 (dichotomy).

Therefore, the 2PL model's α_j and γ_j are

$$\hat{\alpha}_1 = \hat{\alpha}_{12} - \hat{\alpha}_{11} = 0.569 - (-0.569) = 1.138 \text{ and}$$

$$\hat{\gamma}_1 = \hat{\gamma}_{11} - \hat{\gamma}_{12} = -1.152 - 1.152 = -2.304.$$

Of course, Equation 9.4 may be used to obtain item 1's estimated location

$$\hat{\delta}_1 = \hat{\delta}_{1(1,2)} = \frac{\hat{\gamma}_{11} - \hat{\gamma}_{12}}{\hat{\alpha}_{12} - \hat{\alpha}_{11}} = \frac{-2.304}{1.138} = -2.025.$$

For convenience we use the `anova` function to compare the model-level fit of treating item 1 as dichotomous or with four options. As can be seen, treating item 1 as a dichotomy leads to a reduction in BIC from 103,849.8 to 100,533.7. Thus, from a model-data fitting perspective, treating item 1 as a dichotomy shows better fit than when it is treated in polytomous fashion. Of course, as would be expected, there is a “penalty” in terms of item 1’s maximum information for person location estimation. Specifically, reducing the number of categories from four to two leads to a decrease in its maximum information (Figure 9.9).

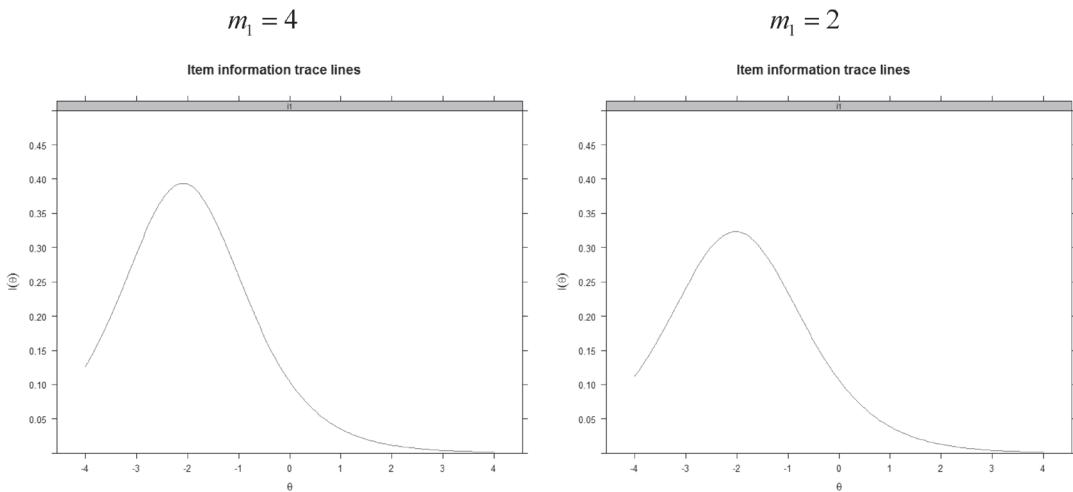


FIGURE 9.9. Item 1 information for $m_1 = 4$ and $m_1 = 2$.

Summary

The nominal response model captures information from nominal polytomous responses. These data may arise in various situations, including, but not limited to, proficiency assessment, attitude, and affective measurement. In addition, the NR model may be applied to dichotomous data and, with the appropriate constraints, to ordinal data. The NR model subsumes the partial credit, generalized partial credit, and two-parameter models.

The NR model contains a slope (i.e., discrimination), α_{jk} , and an intercept, γ_{jk} , parameter for each response category for an item. An alternative parameterization utilizes an overall item discrimination in conjunction with category discrimination parameters. As with previously discussed polytomous models, each response category has a corresponding option response function that describes the probability of selecting a response as a function of θ . Because the slope parameters for an item are constrained to sum to zero across categories, one category always has a monotonically increasing ORF and one a monotonically decreasing ORF. This latter characteristic implies that a particular (e.g., incorrect) response option should be preferred to the remaining responses for individuals with very low θ s. This may be an inappropriate assumption in certain applications. For example, when the NR model is applied to proficiency assessment, the possibility of examinees guessing on items is not addressed. In those cases where guessing may be an issue, then the multiple-choice (MC) or the Bock–Samejima (BS) model may be preferred to the NR model.

The BS model is an extension of the NR model and incorporates a random guessing component. This model states that “don’t know” individuals will guess on the options with the equal probability of $1/m_j$. In contrast, the MC model allows for the possibility that the “don’t know” individuals will guess on the options with potentially unequal

probability. As a result, in addition to the NR model's slope and intercept parameters, the MC model contains a new parameter, ϕ_{jk} , that represents the proportion of respondents who "don't know" and selected category k on item j . Conceptually, the respondents consist of a mixture of individuals who chose a specific option and those who guessed on each of the response options.

The models presented in the previous chapters are applicable to dichotomous, ordered polytomous, and nominal polytomous data. In many respects, all of these models may be viewed as either extensions of the two-parameter model or subsumed by the two-parameter model. As such, we view the two-parameter model as the "nexus model." In this regard, these models share the two-parameter model's assumption that a single latent trait underlies the data. In the next chapter, we extend the two-parameter model to allow for the possibility of multiple latent traits underlying the response data.

Notes

1. For a categorical response variable with m response categories, let $\{p_1, \dots, p_m\}$ contain the corresponding set of response probabilities with $\sum p_k = 1.0$. Given a random sample of N independent cases drawn from a population with $\{p_1, \dots, p_m\}$, then the probability distribution for the sample's observed frequency distribution across the m responses is called a multinomial distribution. (The binomial distribution is a special case of the multinomial distribution when $m = 2$.) For the example we are assuming a multinomial distribution.
2. This is accomplished by using the *multinomial logit model* to predict the observed data (cf. Agresti, 1990). The nominal response model may be considered an example of multinomial logistic regression using the person's latent location and the items' latent characterizations. When $m = 2$ (i.e., binary response data), then we have a single logit equation, and the analysis simplifies to ordinary logistic regression; $\log(p_1/p_0) = \log(p_1) = \gamma + \alpha X$.
3. In general, the transformation matrix, \mathbf{T} , with $(m_j - 1)$ rows and m_j columns contains the following deviation contrasts

$$\mathbf{T} = \begin{bmatrix} -\frac{1}{m_j} & \frac{(m_j - 1)}{m_j} & \dots & -\frac{1}{m_j} & -\frac{1}{m_j} \\ -\frac{1}{m_j} & -\frac{1}{m_j} & \dots & -\frac{1}{m_j} & -\frac{1}{m_j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{m_j} & -\frac{1}{m_j} & \dots & \frac{(m_j - 1)}{m_j} & -\frac{1}{m_j} \\ -\frac{1}{m_j} & -\frac{1}{m_j} & \dots & -\frac{1}{m_j} & \frac{(m_j - 1)}{m_j} \end{bmatrix}.$$

The first column is equal to the negative of the sum of the following $(m_j - 1)$ col-

umns and, as a result, the row sum is 0.0. For example, for a three-response category item ($m_j = 3$) the corresponding \underline{T} would be

$$\underline{T} = \begin{bmatrix} -0.33333 & 0.66667 & -0.33333 \\ -0.33333 & -0.33333 & 0.66667 \end{bmatrix}.$$

Bock's (1972) presentation of \underline{T} has elements whose signs are the opposite of those above. The matrix above corresponds to that used by Thissen et al. (2003).

4. For ordered polytomous data (i.e., graded response data), Samejima (1979) proposed that the probability of a person located at θ responding in category k on item j in the presence of guessing is

$$p_{jk} = \frac{1 - \exp(-\alpha_j(\delta_{j,k+1} - \delta_{j,k}))}{[1 + \exp(-\alpha_j(\theta - \delta_{j,k}))][1 + \exp(\alpha_j(\theta - \delta_{j,k+1}))]} + \frac{1}{m_j [1 + \exp(\alpha_j(\theta - \delta_{j,1}))]} \quad (9.11)$$

where $\alpha_j > 0$, $-\infty < \delta_{j,1} < \delta_{j,2} < \dots < \delta_{j,m_j} < \delta_{j,m_j+1} < \infty$, and $x_j = \{0, 1, \dots, m_j\}$. The model in Equation 9.11 is called Type I Model B by Samejima (1979); her presentation uses $D\alpha_j$ in lieu of our α_j . (Type I Model A is the normal ogive version of Equation 9.11.) For Equation 9.11, the ORF when $k = 0$ is strictly decreasing in θ and represents the “no recognition” category. Therefore, the observed responses are $x_j = \{0, 1, \dots, m_j\}$ with the ORF being strictly increasing in θ when $k = m_j$ and is the item’s correct answer. Both the “ $k = 0$ ” and “ $k = m_j$ ” ORFs have asymptotes of 0 and 1. The ORFs for $x_j = \{1, \dots, (m_j - 1)\}$ are unimodal and asymptotic to 0.

5. The SAS/STAT 14.3 User’s Guide: The IRT Procedure (SAS Institute, 2017) states that `proc irt` will perform an NR model calibration. Earlier versions do not perform an NR model calibration.
6. There is an unscaled χ^2^* and a scaled χ^2^* ($\chi_s^2^*$). In the former case, χ^2^* approximately follows a chi-squared distribution, whereas in the latter case, χ^2^* is rescaled by using a resampling to determine the scaling factor that will transform the values to follow a chi-squared distribution (Stone, 2000). Moreover, the degrees of freedom are adjusted for the use of item parameter estimates in generating the simulated data. Both statistics use resampling for determining the probability of the test statistic. Chapter 5 Endnote 20 mentions that because resampling is used, the p -values will vary across calculations of χ^2^* . As an example, we present the χ^2^* s from five calls to `itemfit`:

```
> itemfit(nrm, fit_stats="X2*")      # first call
   EM cycles terminated after 500 iterations.
   EM cycles terminated after 500 iterations.
   EM cycles terminated after 500 iterations.
      item X2_star p.X2_star
      1    i1    7.841    0.087
      2    i2    4.772    0.848
      3    i3    3.806    0.690
      4    i4    3.878    0.223
      5    i5   12.691    0.734
```

```
> itemfit(nrm,fit _ stats="X2*")      # second call
    item X2 _ star p.X2 _ star
    1   i1    7.841     0.109
    2   i2    4.772     0.820
    3   i3    3.806     0.670
    4   i4    3.878     0.215
    5   i5    12.691    0.712

> itemfit(nrm,fit _ stats="X2*")      # third call
EM cycles terminated after 500 iterations.
    item X2 _ star p.X2 _ star
    1   i1    7.841     0.098
    2   i2    4.772     0.822
    3   i3    3.806     0.693
    4   i4    3.878     0.208
    5   i5    12.691    0.724

> itemfit(nrm,fit _ stats="X2*")      # fourth call
    item X2 _ star p.X2 _ star
    1   i1    7.841     0.091
    2   i2    4.772     0.848
    3   i3    3.806     0.681
    4   i4    3.878     0.234
    5   i5    12.691    0.744

> itemfit(nrm,fit _ stats="X2*")      # fifth call
    item X2 _ star p.X2 _ star
    1   i1    7.841     0.103
    2   i2    4.772     0.857
    3   i3    3.806     0.682
    4   i4    3.878     0.199
    5   i5    12.691    0.703
```

As can be seen, although the χ^2 's remain the same across calls to `itemfit`, their corresponding probabilities vary slightly from call to call. In our example, all tests are nonsignificant at the 5% level. Thus, we conclude that we do not have evidence of item misfit for any of the items.

In addition to the χ_s^{2*} 's presented in Table 9.1, we present four sets of the χ_s^{2*} 's:

```
> itemfit(nrm,fit _ stats="X2*_df")      # second call
    item      X2 _ star _ scaled df.X2 _ star _ scaled RMSEA.X2 _ star _
    scaled p.X2 _ star _ scaled
    1   i1    182.637      154.636      0.004   0.061
    2   i2    140.266      150.906      0.000   0.722
    3   i3    61.982  61.500  0.001   0.459
    4   i4    112.393      94.641   0.004   0.103
    5   i5    121.649      125.146      0.000   0.572

> itemfit(nrm,fit _ stats="X2*_df")      # third call
    item      X2 _ star _ scaled df.X2 _ star _ scaled RMSEA.X2 _ star _
    scaled p.X2 _ star _ scaled
    1   i1    231.634      201.714      0.003   0.073
```

```

2   i2      109.167        118.326        0.000    0.715
3   i3      71.040  70.532  0.001    0.461
4   i4      99.321  84.691  0.003    0.132
5   i5      143.181        147.673        0.000    0.589

> itemfit(nrm, fit_stats="X2*_df")          # fourth call
item   X2_star_scaled df.X2_star_scaled RMSEA.X2_star_
scaled p.X2_star_scaled
1   i1      203.195        174.898        0.003    0.070
2   i2      124.251        135.475        0.000    0.746
3   i3      70.483  70.943  0.000    0.493
4   i4      104.791        88.945  0.003    0.120
5   i5      126.831        132.099        0.000    0.613

> itemfit(nrm, fit_stats="X2*_df")          # fifth call
item   X2_star_scaled df.X2_star_scaled RMSEA.X2_star_
scaled p.X2_star_scaled
1   i1      222.206        190.125        0.003    0.055
2   i2      121.241        130.044        0.000    0.697
3   i3      60.895  61.444  0.000    0.496
4   i4      83.156  69.158  0.004    0.120
5   i5      154.878        159.312        0.000    0.584

```

Because of the resampling strategy used in calculating χ_s^{2*} , both its value and p -value vary across calls. Again, because all tests are nonsignificant at the 5% level we conclude that we do not have evidence of item misfit for any of the items. This conclusion is consistent with our items' RMSEAs. Note: Calculating χ^2 and χ_s^{2*} takes several minutes. It is recommended that whenever an item's $p(\chi_s^{2*})$ is close to the significance level (e.g., item 1) one repeatedly recalculate χ_s^{2*} to discern the consistency in the p -values (e.g., 95% of the p -values are nonsignificant) before making a decision about rejecting the null hypothesis. This would also be the recommendation when $p(\chi^{2*})$ is close to the significance level.

7. To obtain the item parameter estimates in Thissen et al.'s (2010) parameterization, we set the `IRTpars` argument to FALSE in our call to `coef`:

```

> coef(nrm, simplify=T, IRTpars=F)
$items
  a1 ak0  ak1  ak2  ak3 d0      d1      d2      d3
i1 0.671  0 1.973 0.052  1 0  4.643  1.780  1.196
i2 1.303  3 1.822 1.550  2 0 -2.821 -3.561 -2.624
i3 0.429  3 0.610 0.723  2 0 -2.997 -3.074 -1.469
i4 0.484  1 0.644 1.867  3 0 -0.426  0.399  1.845
i5 0.961  2 1.010 0.931  3 0 -2.484 -2.534  2.423
:

```

These are the estimates that are transformed to obtain the Bock parameterization values. Equation 9.6's overall discrimination parameter (α_j) is `a1`, the `ak0`, `ak1`, ..., `ak3` correspond to the $\alpha_{j,k+1}^s$ s, and the `ds` correspond to the $\gamma_{j,k+1}$ s. One sees the identification constraints in the values for `ak0` and `d0`.

To convert these estimates to the Bock parameterization, one calculates the multiplicative discrimination $\hat{\alpha}_j \hat{\alpha}_{j,k+1}^s$ for each category for each item and then centers them itemwise. Using item 2 as an example, we have for category 1 $\hat{\alpha}_{21} = \hat{\alpha}_2 \hat{\alpha}_{2,0}^s = 1.303 * 3 = 3.909$. The remaining values are:

item	$\hat{\alpha}_j \hat{\alpha}_{j1}$	$\hat{\alpha}_j \hat{\alpha}_{j2}$	$\hat{\alpha}_j \hat{\alpha}_{j3}$	$\hat{\alpha}_j \hat{\alpha}_{j4}$	mean
i1	0.000	1.324	0.035	0.671	0.507
i2	3.909	2.374	2.020	2.606	2.727
i3	1.287	0.262	0.310	0.858	0.679
i4	0.484	0.312	0.904	1.452	0.788
i5	1.922	0.971	0.895	2.883	1.668

Upon centering (e.g., for item 2: $3.909 - 2.727 = 1.182$), we obtain Bock's parameterization:

item	$\hat{\alpha}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\alpha}_{j3}$	$\hat{\alpha}_{j4}$
i1	-0.507	0.816	-0.473	0.164
i2	1.182	-0.353	-0.708	-0.121
i3	0.608	-0.418	-0.369	0.179
i4	-0.304	-0.476	0.116	0.664
i5	0.254	-0.697	-0.773	1.215

Within rounding, these values match those shown in Table 9.1.

To obtain our intercepts, we simply center each item's ds. For convenience we redisplay each item's ds with their corresponding mean:

item	d0	d1	d2	d3	mean
i1	0	4.643	1.780	1.196	1.905
i2	0	-2.821	-3.561	-2.624	-2.252
i3	0	-2.997	-3.074	-1.469	-1.885
i4	0	-0.426	0.399	1.845	0.455
i5	0	-2.484	-2.534	2.423	-0.649

As an example, for item 1 we have $\hat{\gamma}_{11} = d0 - \text{Avg}(d) = 0 - 1.905 = -1.905$. Our remaining values are:

item	$\hat{\gamma}_{j1}$	$\hat{\gamma}_{j2}$	$\hat{\gamma}_{j3}$	$\hat{\gamma}_{j4}$
i1	-1.905	2.738	-0.125	-0.709
i2	2.252	-0.570	-1.310	-0.373
i3	1.885	-1.112	-1.189	0.416
i4	-0.455	-0.881	-0.056	1.391
i5	0.649	-1.835	-1.885	3.072

These values agree (within rounding) with the corresponding values in Table 9.1.

8. mirt version 1.30 does not produce standard errors for the item parameter estimates.
9. We call mirt specifying using the NR model for each item to demonstrate that when NR model is applied to a dichotomous item, the NR model is equivalent to the 2PL model. An alternative and general approach is to specify the model to use for each item. In this fashion the mixed model calibration is evident. For instance, as

arguments to `itemtype`, we specify 2PL for item 1 and the nominal model for the remaining four items. We could write out nominal four times or use the replicate function (`rep`):

```
> print((nrmilm2 = mirt(data = GenSciencedatailm2, model = freqbasis,
+ itemtype = c("2PL", rep("nominal", 4)), SE = T)))
Iteration: 32, Log-Lik: -50142.210, Max-Change: 0.00010

Calculating information matrix . . .
Call:
mirt(data = GenSciencedatailm2, model = freqbasis, item-
type = c("2PL",
rep("nominal", 4)), SE = T)

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 32 EM iterations.
mirt version: 1.30
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Information matrix estimated with method: Oakes
Condition number of information matrix = 475.6606
Second-order test: model is a possible local maximum
Log-likelihood = -50142.21
Estimated parameters: 26
AIC = 100336.4; AICc = 100336.5
BIC = 100533.7; SABIC = 100451.1
G2 (485) = 633.09, p = 0
RMSEA = 0.005, CFI = NaN, TLI = NaN
> coef(nrmilm2, simplify = T, IRTpars = T)
$items
     a      b      g      u      a1      a2      a3      a4      c1      c2      c3      c4
i1 1.137 -2.027  0 1    NA      NA      NA      NA      NA      NA      NA      NA
i2  NA     NA     NA     NA   1.180 -0.362 -0.710 -0.108  2.251 -0.577 -1.312 -0.363
i3  NA     NA     NA     NA   0.608 -0.424 -0.371  0.187  1.887 -1.116 -1.190  0.419
i4  NA     NA     NA     NA  -0.299 -0.474  0.113  0.660 -0.452 -0.879 -0.057  1.389
i5  NA     NA     NA     NA   0.254 -0.693 -0.777  1.216  0.647 -1.829 -1.890  3.072

$means
F
0
$cov
F
F 1
```

As another example, say the fifth item is to be fitted with the GR model, our call to mirt is: `mirt(data = GenSciencedatai1m2, model = freqbasis, itemtype = c("2PL", rep("nominal", 3), "graded"), SE = T))`.

10. With Thissen et al.'s (2010) scoring function parameterization, item 1's overall discrimination parameter estimate a_1 corresponds to $\hat{\alpha}_1$ and its d_1 is $\hat{\gamma}_1$. Therefore, $\hat{\delta}_1 = d_1/a_1$.

10

Models for Multidimensional Data

The models presented in previous chapters contained a single parameter, θ , to reflect a person's location on a continuous latent variable. That is, these models were predicated on a unidimensional latent space. However, in some situations it may be more realistic to hypothesize that a person's response to an item is due to their locations on multiple latent variables. As such, we have a *multidimensional* latent space. In this chapter we present models that use multiple person location parameters to describe an individual's response behavior.

Conceptual Development of a Multidimensional IRT Model

One might encounter multidimensionality in various situations. For instance, consider an instrument designed to measure self-efficacy in overcoming barriers to healthy eating. If, theoretically, healthy eating self-efficacy involves cognitive and affective dimensions, then responses to this instrument are a function of a respondent's locations on these dimensions. A second example of multidimensionality is performance on a mathematical word problem. In this case, we may have one dimension that reflects mathematics proficiency and another, reading proficiency.

These two examples describe two possible multidimensional scenarios that differ from one another in how the latent person variables interact to produce the observed responses. In the second example, an individual with highly developed reading proficiency might be able to compensate, to some extent, for their lower mathematics proficiency to correctly respond to a mathematical word problem. In contrast, with our self-efficacy example a respondent's location on the cognitive dimension could not compensate for their location on the affective dimension. In short, the mathematical word problem scenario reflects a compensatory multidimensional situation where a person's location(s) on one or more dimension(s) can compensate for their location(s) on other latent variable(s). Conversely, the self-efficacy example presents a noncompensatory or partially compensatory multidimensional case. With a noncompensatory

situation, a person's location on one or more latent variables do not compensate for their location(s) on other latent variable(s). Therefore, modeling these types of data requires distinguishing between compensatory and noncompensatory multidimensional situations.

Models for noncompensatory multidimensional scenarios are called *noncompensatory* (a.k.a., *partially compensatory*) models, whereas models for compensatory multidimensional scenarios are referred to as *compensatory* multidimensional models. Examples of noncompensatory models may be found in Sympson (1978) and Whitley (1980). These models have not seen as much attention as the compensatory models, in part, because of estimation difficulties; also see Spray, Davey, Reckase, Ackerman, and Carlson (1990).¹ In the following we focus on two common compensatory models.

To develop a *multidimensional item response theory* (MIRT) model we assume for simplicity, but without loss of generalizability, that we are interested in modeling a two-dimensional latent space using binary data. It is possible to generalize our dichotomous model to polytomous responses. For instance, we have the multidimensional-graded response (MGR; de Ayala, 1994) generalized partial credit (MGPC; Yao & Schwarz, 2006), or nominal response (MNR; Revuelta, 2014) models. Of course, we can have more than two dimensions although graphical depiction becomes problematic with more than two dimensions.

From Chapters 2 and 5 we know that with dichotomous unidimensional models the probability of a response of 1 is a function of the unweighted or weighted distance between item j 's location and person i 's location (e.g., $\alpha_j(\theta_1 - \delta_j)$). This idea may be extended to the two-dimensional latent space. At the simplest level, the probability of a response of 1 is a function of the difference between item j 's location and person i 's location on dimension 1 (e.g., $(\theta_{i1} - \delta_{j1})$) and the difference between the item's location and the person's location on dimension 2 (e.g., $(\theta_{i2} - \delta_{j2})$); the second subscript indicates the dimension. However, simply using $(\theta_{i1} - \delta_{j1})$ and $(\theta_{i2} - \delta_{j2})$ in the logistic function implies that the item's relationship to each dimension is the same for both dimensions. This may not be the case in all situations. For instance, assume that an examination consists of algebra word problems. One question may require greater knowledge of algebra than it does reading proficiency. As a result, the algebra dimension is a stronger determinant of a response of 1 on this item than is the reading proficiency dimension, although both are required to answer the question. In contrast, to be correctly answered, another item may require less algebra knowledge than it does reading proficiency. Therefore, to reflect that an item's relationship to a dimension may vary across dimensions, these logits must be weighted to represent these item–dimension relationships.

A mechanism for weighting the logits for a dimension comes from factor analysis. Recall that in factor analysis the item's loading on a factor reflects the relationship between the item and the dimension. As presented in Appendix C, an item's loading is related to its discrimination parameter. Consequently, using the item's discrimination provides a mechanism to capture the item's relationships to the underlying dimensions. Specifically, an item's discrimination on each dimension can serve as a weight for the logit for that dimension. Incorporating these ideas into the logistic function produces

$$p(x_{ij} = 1 | \theta_{i1}, \theta_{i2}, \alpha_{j1}, \alpha_{j2}, \delta_{j1}, \delta_{j2}) = \frac{e^{\alpha_{j1}(\theta_{i1}-\delta_{j1})+\alpha_{j2}(\theta_{i2}-\delta_{j2})}}{1+e^{\alpha_{j1}(\theta_{i1}-\delta_{j1})+\alpha_{j2}(\theta_{i2}-\delta_{j2})}}, \quad (10.1)$$

where α_{j1} and α_{j2} are item j 's discrimination parameters on dimension 1 and 2, respectively. The model in Equation 10.1 states that the probability of a response of 1 is a function of the distance between person i and item j 's location on each dimension and the item's relationship to each dimension. Equation 10.1 may be seen as a generalization of the 2PL model to a two-dimensional latent space.^{2,3} As with the 2PL model, items may vary in their discrimination parameters.

Alternatively, Equation 10.1 may be reparameterized into a slope–intercept form by introducing an intercept parameter, γ_j . Recall that in the unidimensional case this parameter reflects the interaction between an item's location and its capacity to discriminate among individuals (e.g., Chapter 2, Equation 2.3). Applying this idea to the two-dimensional case, Equation 10.1 becomes

$$p(x_{ij} = 1 | \theta_{i1}, \theta_{i2}, \alpha_{j1}, \alpha_{j2}, \gamma_j) = \frac{e^{(\alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2}) + \gamma_j}}{1+e^{(\alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2}) + \gamma_j}}, \quad (10.2)$$

where

$$\gamma_j = -(\alpha_{j1}\delta_{j1} + \alpha_{j2}\delta_{j2}) = -\sum_{f=1}^2 \alpha_{jf}\delta_{jf}$$

One may generalize from the two-dimensional situation to F latent variables or dimensions. This generalization of Equation 10.2 yields the (compensatory) *multidimensional two-parameter logistic* (M2PL) model (McKinley & Reckase, 1983a; Reckase, 1985, 2009)

$$p(x_{ij} = 1 | \underline{\theta}_i, \underline{\alpha}_j, \gamma_j) = \frac{e^{\sum \alpha_{jf}\delta_{jf} + \delta_j}}{1+e^{\sum \alpha_{jf}\delta_{jf} + \delta_j}} = \frac{e^{\underline{\alpha}'_j \underline{\theta}_i + \delta_j}}{1+e^{\underline{\alpha}'_j \underline{\theta}_i + \delta_j}} \quad (10.3)$$

where $p(x_{ij} = 1 | \underline{\theta}_i, \underline{\alpha}_j, \gamma_j)$ is the probability of a response of 1 on item j by person i , given their locations on each of the F -dimensions and the item characteristics of $\underline{\alpha}_j$ and γ_j . The (column) vector $\underline{\theta}_i$ contains person i 's location parameters on each of the F -dimensions (i.e., $\underline{\theta}'_i = (\theta_{i1}, \dots, \theta_{if}, \dots, \theta_{iF})$). The item is characterized by a vector, $\underline{\alpha}_j$, containing item j 's discrimination parameters on each of the F -dimensions (i.e., $\underline{\alpha}'_j = (\alpha_{j1}, \dots, \alpha_{jf}, \dots, \alpha_{jF})$) and an intercept parameter, γ_j ; the prime symbol on $\underline{\alpha}_j$ indicates that $\underline{\alpha}'_j$ and $\underline{\theta}'_i$ indicate row vectors. The discrimination parameters indicate the item's sensitivity to differences in person locations in the latent space in a particular direction (e.g., along the θ_1 -axis). The intercept, γ_j , reflects the interaction of the item's location and discrimination parameters. In a proficiency assessment situation γ_j would be interpreted as *related* to an item's difficulty/easiness. Analogous to Equation 2.3, the intercept parameter in the F -dimensional case is given by

$$\gamma_j = -\sum_{f=1}^F \alpha_{jf}\delta_{jf} \quad (10.4)$$

The number of elements in $\underline{\theta}_i$, and therefore α_j , equals the number of interpretable dimensions underlying the data. In the following, we use p_j in lieu of $p(x_{ij} = 1 | \underline{\theta}_i, \alpha_j, \gamma_j)$.

Graphically, Equation 10.3 (or Equations 10.1 or 10.2) would produce a sigmoidal-shaped analog to the IRF. This analog is called the *item response surface* (IRS). For a two-dimensional situation (i.e., $F = 2$) we need three dimensions to display the IRS where two of the axes represent the two latent person variables and the third axis reflects the probability of a response of 1.

Figure 10.1 contains an example of an IRS for the two-dimensional case where $\alpha_{j1} = 2.0$, $\alpha_{j2} = 0.5$, $\delta_{j1} = -1.25$, and $\delta_{j2} = 1$ (or $\gamma_j = 2$). We see that the probability of a response of 1 increases as θ increases along each dimension such that the surface is asymptotic at 1. Conversely, as θ decreases, we see that the surface becomes asymptotic with $p_j = 0$ (the left foreground). Moreover, we can see the compensatory nature of the model at work. For instance, even when a person is located at -4 on dimension 2 ($\theta_2 = -4$), if they're located high enough on dimension 1 (e.g., $\theta_1 = 4$), the probability of a response of 1 is greater than 0.98.

Examining the surface shows that the IRS has different (conditional) slopes. This observation is easy to demonstrate by taking "slices" out of the IRS. For instance, for a person located at $\theta_1^* = -1.2$, our slice through the IRS results in a conditional trace line along θ_2 (i.e., a line from the right foreground to the left background when $\theta_1^* = -1.2$).

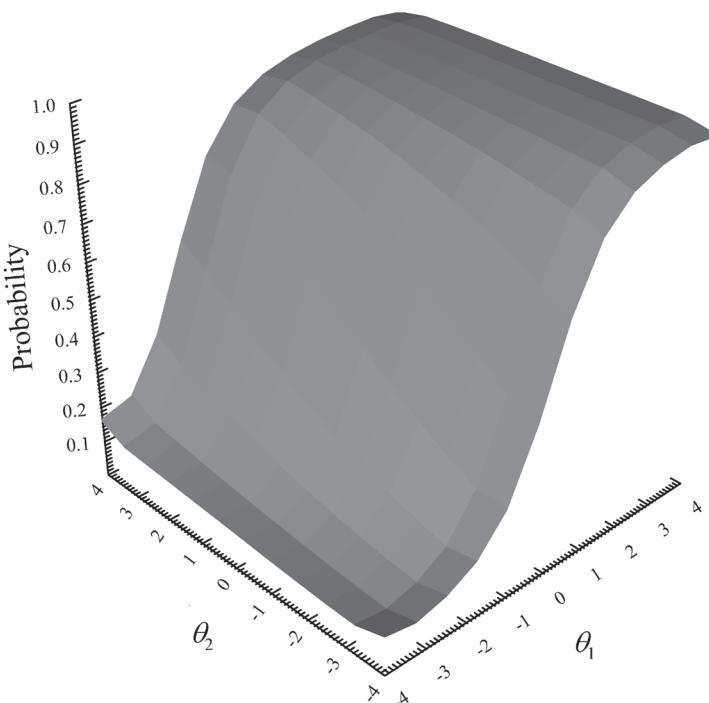


FIGURE 10.1. Item response surface for a two-dimensional item ($\alpha_{j1} = 2.0$, $\alpha_{j2} = 0.5$, and $\gamma_j = 2.0$).

We can take a second conditional trace line by slicing through the IRS along dimension 2 for $\theta_2^* = -1.8$ (i.e., a line from the left foreground to the right background when $\theta_2^* = -1.8$). These two conditional trace lines are shown in Figure 10.2. Our $\theta_2^* = -1.8$ conditional trace line has a substantially greater slope than does the $\theta_1^* = -1.2$ conditional trace line. These different slopes reflect that, in part, the item's relationship to dimension 1 is different from its relationship to dimension 2.

With the 2PL model, the probability of a response of 1 on item j equals 0.5 whenever $\theta_i = \delta_j$ (i.e., the logit equals 0). Similarly, with the M2PL model whenever the logit equals 0, then the probability of a response of 1 on item j equals 0.5. However, across the two dimensions, there are multiple combinations of θ s that for a given item parameter set result in the logit equaling 0.0. As a result, and unlike the unidimensional case with a single point for $p_j = 0.5$, in the multidimensional situation there is a line of points for which $p_j = 0.5$. We refer to this line of points as the *inflection line*. This property is easily shown using a contour plot to represent the IRS.

With a contour plot, the contour lines represent points of equal probability. Figure 10.3 contains the contour plot corresponding to the IRS shown in Figure 10.1. As indicated by the legend, each contour line shows a different level of p_j as well as the various combinations of θ s that interact with the item's parameters to produce these different p_j s. The pattern seen is that the probabilities increase as one moves from left to right in the graph. For example, the rightmost line traces the various combinations of θ_1 and

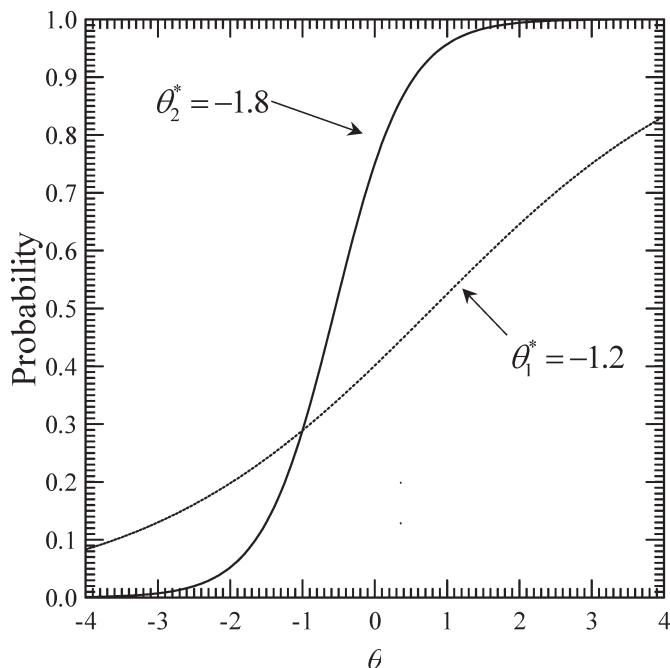


FIGURE 10.2. Conditional trace lines of the IRS shown in Figure 10.1.

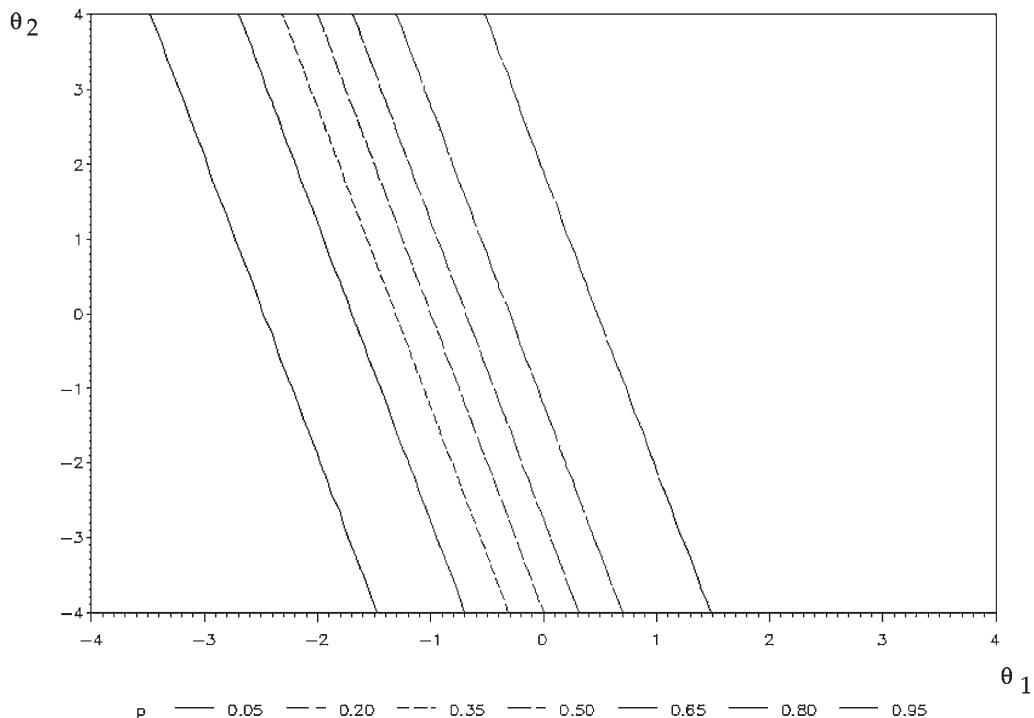


FIGURE 10.3. Contour plot of IRS in Figure 10.1 ($\alpha_{j1} = 2.0$, $\alpha_{j2} = 0.5$, and $\gamma_j = 2.0$).

θ_2 , which given this item's parameters, results in a $p_j = 0.95$. The fourth contour line from the left reflects a p_j of 0.5 and is the IRS's inflexion line. The steepness of an IRS is represented by the proximity of the contour lines. The steeper the IRS, the closer the corresponding contour lines are to one another. Conversely, contour lines that are comparatively farther apart reflect a less steep portion of the IRS. This item shows a relatively steep (discriminating) portion in the center of the band of contour lines.

A logit space plot helps to discuss the intercept parameter in the multidimensional case. Figure 10.4 shows the logit space plot for the item shown in Figures 10.1 and 10.3. Rather than the logit regression line seen with the unidimensional models (e.g., Chapters 2 and 9), we now have a *logit regression plane*. This plane is analogous to the regression plane for a two-predictor multiple regression model. (To facilitate interpreting the ordinate, the axes have been truncated to begin at 0.0 and the θ_2 axis has been reversed from the way it is shown in Figure 10.1.) We see that not only does the plane have a steeper slope along dimension 1 than along dimension 2, reflecting that α_{j1} is four times larger than α_{j2} , but that the plane intersects the ordinate at $\gamma_j = 2$. Given that α_{j1} is larger than α_{j2} , we know that this item is better at discriminating among respondents on θ_1 than on dimension 2. One implication of this is that this item provides greater information for locating individuals along dimension 1 than along dimension 2.

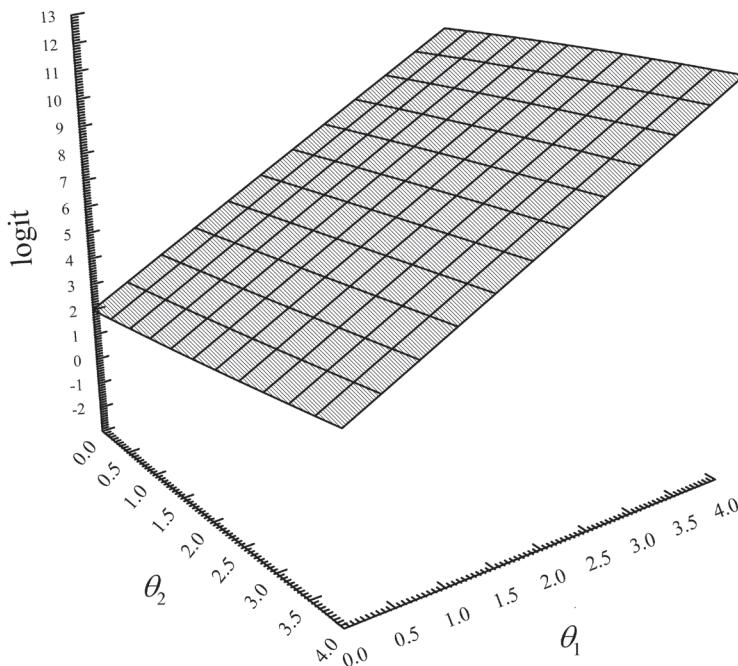


FIGURE 10.4. Multivariate logit space plot of the item in Figure 10.1 ($\alpha'_j = (2.0, 0.5)$ and $\gamma_j = 2.0$).

Multidimensional Item Location and Discrimination

Because γ_j involves both discrimination and location parameters, we cannot interpret γ_j as a location parameter. However, we can calculate an index, the *multidimensional item location* (Δ_j), that can be interpreted as a location parameter. In an analogous fashion, we can determine an item's capacity to discriminate among individuals across all dimensions. This index is item j 's *multidimensional item discrimination* parameter, A_j . Consequently, although an item may not be a pure measure of a single latent dimension, A_j and Δ_j provide a form of data reduction of the item's multidimensional characterizations. That is, regardless of the number of dimensions, A_j provides a single (i.e., scalar) value that represents the best that item j can discriminate across all the dimensions. Similarly, the item's location in the multidimensional space is given by a single value, its Δ_j , and its direction cosine(s). The calculation of an item's A_j is informed by its Δ_j , and the calculation of Δ_j uses A_j . In the following, we begin with the concept of a multidimensional item location and then discuss an item's multidimensional discrimination parameter.

Item j 's multidimensional item location is given by

$$\Delta_j = \frac{-\gamma_j}{A_j}, \quad (10.5)$$

where all terms are defined above. An item's multidimensional item location indicates

the *distance* from the origin in the θ space to the point of maximum slope (i.e., the item's maximum discrimination) in a *particular direction* from the origin; Equation 10.5 is also called the item's *multidimensional item difficulty*.⁴ This definition is based on the assumption that the most reasonable point to use in defining item j 's Δ_j is the point where the item is most discriminating (Reckase, 1985, 2009).

Ignoring the directionality issue for now, Δ_j 's definition is analogous to how the unidimensional item location is defined. That is, δ_j is the point where the IRF has its maximum slope—its inflection point. A negative sign on Δ_j corresponds to a unidimensional model's negative location parameter, whereas a positive Δ_j reflects a unidimensional model's positive δ_j . Moreover, because A_j represents the item's discrimination capacity, Equation 10.5 is analogous to defining an item's unidimensional location in terms of its intercept and discrimination parameters (cf. Equation 2.5 in Chapter 2). Accordingly, if we rearrange Equation 10.5 we can obtain an expression of the intercept in terms of the multidimensional item location and multidimensional discrimination parameters

$$\gamma_j = -A_j \Delta_j. \quad (10.6)$$

Equation 10.6 is analogous in form to the intercept's definition with, say, the 2PL model (i.e., $\gamma_j = -\alpha_j \delta_j$).

As mentioned and demonstrated above, an IRS has many slopes. That is, the slope of the IRS at a particular point may vary from that at another point, depending on where we are on the surface. This is easily seen in Figure 10.2 where the conditional trace lines reflect two different directions along the IRS. As we see, the slopes of the conditional trace lines at the intersection point differ from one another. Implied in Figures 10.1 and 10.2 is that the slope reflects a change in the surface for a unit change in the $\theta_1-\theta_2$ plane in a particular direction. To determine this direction, we need a reference or starting point for the path that we take across the surface and along which we calculate the unit change. Stated another way, there are many different starting points we can proceed from as we traverse the surface with the slope along and across each of these paths varying. By convention, the origin of the space serves as the reference point of interest (i.e., the point defined by $\theta_1 = \dots = \theta_F = 0$; Reckase, 1985). From the origin we can determine the IRS's maximum slope in multiple directions. It is the maximum value of these maxima that is used in defining item j 's ω_{Δ_j} .

Conceptually, to determine the IRS's point of maximum slope we first determine the slopes' values in different directions. Once we know these values, we can determine where the slopes are maximized and the direction of the maximum with respect to the origin. Let ω_{jf} represent a direction or angle from the f th dimension to a point of maximum slope for item j . For the M2PL model, the *value* of the slope in the direction given by ω_{jf} is (Reckase, 1985)

$$p_j(1-p_j) \sum_{f=1}^F \alpha_{jf} \cos \omega_{jf}. \quad (10.7)$$

Given that at the inflection line (i.e., $p_j = 0.5$) the slope in the direction given by ω_{if} is at its maximum, then the slope's value at a point on the inflection line equals

$$0.25 \sum_{f=1}^F \alpha_{jf} \cos \omega_{jf}^\circ. \quad (10.8)$$

Hypothetically, if we apply Equation 10.8 to angles from 0° to 90° with respect to the f^{th} dimension, we obtain a series of corresponding slope values that represent the maximum at that angle (i.e., direction). We are interested in the maximum of these values and its direction. For instance, for our example item with $\underline{\alpha}' = (2.0, 0.5)$ and the angles from 12° to 16° (with respect to the dimension θ_1) in one degree increments, the application of Equation 10.8 gives us the corresponding slope values of $(0.51506, 0.51530, 0.51539, 0.51532, 0.51509)$; that is, the slope at 12° is 0.51506, at 13° it is 0.51530, and so on. The maximum (slope) of these maxima is 0.51539. We now present an analytical approach alternative to this brute force strategy.

To determine the location of the maximum value of the series of slopes given by Equation 10.8, we need to identify its direction. The direction of the maximum slope from the origin is

$$\cos \omega_{if}^\circ = \frac{\alpha_{if}}{\sqrt{\sum_{f=1}^F \alpha_{jf}^2}} \quad (10.9)$$

or, alternatively, the angle with respect to dimension f is

$$\omega_{if}^\circ = \arccos \left(\frac{\alpha_{if}}{\sqrt{\sum_{f=1}^F \alpha_{jf}^2}} \right). \quad (10.10)$$

(In vector geometry, Equation 10.9 is sometimes referred to as a direction cosine.) As an example, given $\underline{\alpha}' = (2.0, 0.5)$, the maximum slope of 0.51539 occurs at an angle of approximately 14° with respect to θ_1 (below we show the calculation to obtain the angle). Therefore, the distance from the origin to the maximum slope of 0.51539 in the direction of 14° with respect to θ_1 is the item's Δ_j . We determine this distance below.

As mentioned above, the value of the slope at the inflection point in a direction given by ω_{if} is obtained by Equation 10.8. Equation 10.9 tells the specific direction to find the maximum slope of the maxima slopes. By substituting Equation 10.9 into Equation 10.8, the maximum slope is

$$0.25 \sum_{f=1}^F \alpha_{jf} \cos \omega_{jf}^\circ = 0.25 \sum_{f=1}^F \alpha_{jf} \left(\frac{\alpha_{jf}}{\sqrt{\sum_{f=1}^F \alpha_{jf}^2}} \right) = 0.25 \sqrt{\sum_{f=1}^F \alpha_{jf}^2} \quad (10.11)$$

Equation 10.11 directly gives the maximum slope of the maxima obtained through Equation 10.8. For example, applying Equation 10.11 to our example item yields, as identified above, that the maximum slope is 0.51539. Equation 10.11 shows that, analogous to the 2PL model, the item's α_{if} s are related to the maximum slope at the inflexion line (McKinley & Reckase, 1983a; Reckase, 1985).

Because we now know the direction in which to proceed to calculate Δ_j , we can discuss the other aspect of Δ_j 's definition, the item's multidimensional discrimination capacity, A_j . We can determine item j 's capacity to discriminate individuals across all F-dimensions by calculating its multidimensional item discrimination parameter (Reckase & McKinley, 1991)⁵

$$A_j = \sqrt{\sum_{f=1}^F \alpha_{if}^2} \quad (10.12)$$

Reckase and McKinley (1991) define A_j to be a function of the slope of the IRS defined by the model at the steepest point in the direction indicated by the multidimensional item location, Δ_j . As stated above, the slope at the "steepest point" is the maximum of the slope maxima and occurs on the IRS's inflexion line. The larger the value of A_j , the greater item j 's discrimination capacity across the F-dimensions. For an item that measures only one dimension, A_j reduces to the unidimensional α_j because the α_{if} s for the other dimensions would be equal to 0. As is the case with the α_{if} s, A_j is equal to four times the maximum slope.

For our example item (Figures 10.1 and 10.3), its multidimensional discrimination would be

$$A_j = \sqrt{\sum_{f=1}^F \alpha_{if}^2} = \sqrt{2.0^2 + 0.5^2} = 2.062.$$

Given that the item's $\gamma_j = 2$, then this item's multidimensional item location is

$$\Delta_j = \frac{-\gamma_j}{A_j} = \frac{-2.0}{2.062} = -0.9701$$

in a direction given by Equation 10.10 of

$$\omega_{il}^\circ = \arccos \left(\frac{\alpha_{j1}}{\sqrt{\sum_{f=1}^F \alpha_{if}^2}} \right) = \arccos \left(\frac{2.0}{2.062} \right) = 14.09^\circ$$

with respect to dimension 1. Alternatively, with respect to dimension 2, the direction of maximum slope is

$$\omega_{j2}^\circ = \arccos \left(\frac{\alpha_{j2}}{\sqrt{\sum_{f=1}^F \alpha_{jf}^2}} \right) \frac{\alpha_{j2}}{\sqrt{\sum_{f=1}^F \alpha_{jf}^2}} = \frac{0.5}{2.062} = 75.97^\circ.$$

Therefore, if we proceed from the origin in a direction of 14.09° from dimension 1 a distance of -0.9701 logits, then we arrive at the item's point of maximum slope; the slope's value at this point is 0.51539. This angle's magnitude indicates this item primarily measures θ_1 . If this is a proficiency item, then given its Δ_j it might be considered to be somewhat easy and reasonably discriminating given its A_j .

The A_j of 2.062 is the maximum discrimination capacity on the inflexion line. This can be shown by applying Equation 10.8 and multiplying the corresponding slope values by 4. (In the following, all angles are with respect to the θ_1 dimension.) For example, if we increase the angle to 45° , then the item's discrimination capacity decreases from 2.062 to 1.7678. If we continue until the angle is 90° , then the item's discrimination capacity would decrease further to 0.50 (i.e., α_{j2} 's value). Conversely, if our angle is 0° (i.e., the θ_1 axis), then the item's discrimination capacity is ($\alpha_{j1} =$) 2.0.

It should be noted that we should compare items' A_j s only when they are measuring in the same direction (cf. Reckase & McKinley, 1991). To compare the discrimination capacity of items that are measuring in different directions, we need to select a common direction ω and then calculate the *directional discrimination*, $A_{\omega j}$, for an item j

$$A_{\omega j} = \sum_{f=1}^F \alpha_{jf} \cos \omega_{jf}^\circ \quad (10.13)$$

Item Vectors and Vector Graphs

Although IRS depictions (e.g., Figure 10.1) are useful, they can become cumbersome when one wants to simultaneously present multiple items. By using the traditional multivariate graphical technique of a vector graph, we can represent the discrimination(s) and location(s) of one or more items. This type of graph allows us to simultaneously present not only an item's location (Δ_j) and how well it discriminates (A_j), but also which dimension, if any, it measures best. Historically, multivariate techniques such as factor analysis have used vector graphs to represent how well items load on different factors (e.g., see Thurstone, 1947). Each item is represented by a vector. (A vector is typically represented by an arrow that has a starting point, a specified length, and points in a particular direction.)

To graph an item vector, we need to know its starting point, its length, and its direction in the multidimensional space. The item vector's starting point is given by Δ_j (i.e., the distance from the θ space's origin to the point of maximum slope) and its length by A_j .⁶ In general, the direction of the vector with respect to a particular dimension f is given by Equation 10.10. In the following we use the horizontal axis to represent θ_1 , which also serves as our reference axis.

The angle ω (in degrees) from the reference axis to the vector can be calculated from A_j by substituting Equation 10.12 into Equation 10.10 to obtain

$$\omega_{j1}^\circ = \arccos\left(\frac{\alpha_{j1}}{A_j}\right). \quad (10.14)$$

Once we know an item's A_j , Δ_j , and ω_{ij} , then we have all the components necessary to represent the item as a vector. For example, the item shown in Figure 10.1 has an angle of

$$\omega_{j1}^\circ = \arccos\left(\frac{\alpha_{j1}}{A_j}\right) = \arccos\left(\frac{2.0}{2.062}\right) = 14.09^\circ$$

with respect to the reference axis, θ_1 ; above we said this angle was approximately 14° . Therefore, if for this item we proceed from the origin a distance of $\Delta_j = -0.9701$ at an angle of $\omega_{j1}^\circ = 14.09^\circ$, we come to the item's location as well as the IRS's point of maximum slope. If we proceed an additional distance of $A_j = 2.062$, then we have traversed the item's multidimensional discrimination capacity.

In Figure 10.5 we graphically represent how the relationships among Δ_j , A_j , and ω_j define an item vector. Before discussing the item vectors, we describe the figure's layout. The graph uses the standard quadrant notation with the quadrants labeled in their

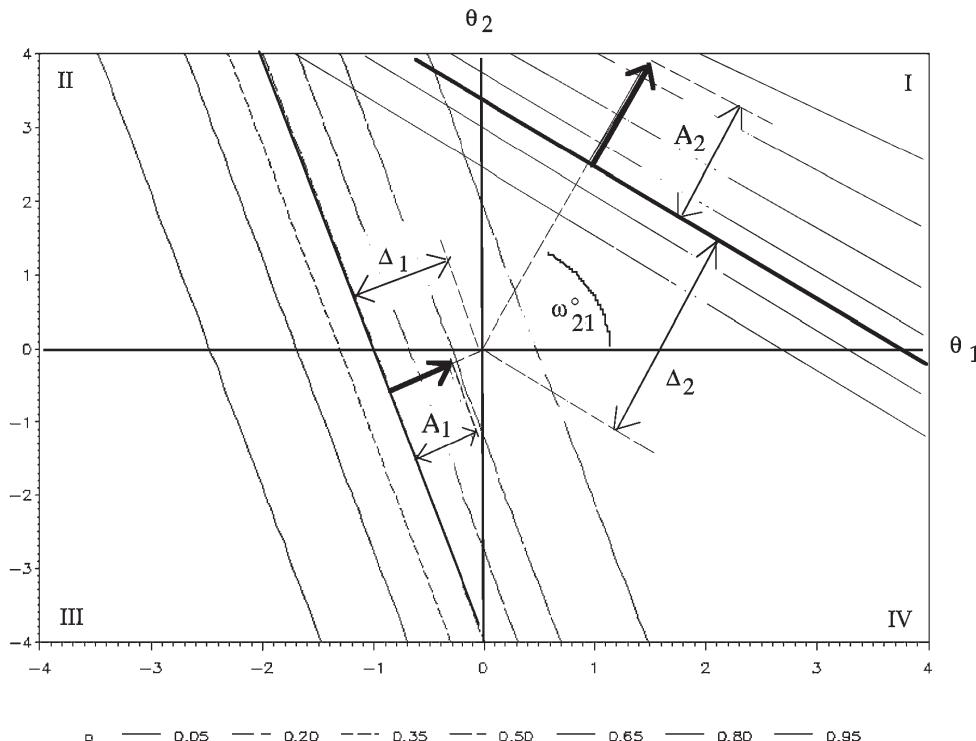


FIGURE 10.5. Contour and vector plot for two items with different Δ_j s, A_j s, and ω_j s.

respective outside corners. The point of origin for this two-dimensional space is in the center of the graph where all four quadrants meet and the θ_1 and θ_2 continua intersect; the metrics for θ_1 and θ_2 are on the bottom and left margins of the graph, respectively. Item vectors representing two items are presented as bold arrows; the double-headed arrows indicate distances. The contour plot for item 1 primarily occupies quadrants II and III, whereas item 2's contour plot primarily occupies quadrant I. For each contour plot the IRS's inflexion line is in bold. Although we have overlaid the items' contour plots to help clarify the relationship between an item's vector representation and its Δ_j , A_j , and ω_j values, vector plots do not normally have contour plots embedded in them (e.g., see Figure 10.8). In a proficiency assessment situation, items in the top right corner of quadrant I would be considered to be difficult items and those in the bottom left corner of quadrant III would be considered to be easy items.

The item vector for item 2 appears in quadrant I (i.e., the upper right part of the figure). Recall that with the 2PL model the item's location is defined at the IRF's point of the inflexion. Analogously, with the M2PL model the multidimensional item location is defined at the IRS's line of inflexion. If we project from the origin to the closest point on the line of inflexion, then this distance is item 2's multidimensional item location, Δ_2 . As we see, the item vector begins on the item's inflexion line (i.e., where $p_j = 0.5$). The angle between this projection and our reference axis is ω_{21}° . The length of item 2's vector is the item's multidimensional discrimination capacity, A_2 , and reflects how well the item discriminates across the F-dimensional space. This depiction shows that item 2 is closer to θ_2 than it is to θ_1 . Consequently, item 2 is primarily measuring θ_2 and discriminating best along this dimension. (If the item had measured *only* θ_2 , then ω_{21}° would have been 90° , and if it had measured *only* θ_1 , then ω_{21}° would have been 0° .)

Item 1 is shown as falling in quadrant III. As is the case with item 2, the item vector's starting point is on the inflexion line. The perpendicular distance from this line to the origin is its Δ_1 and the length of the vector is its A_1 . (We do not show ω_{11}° , but it could be depicted as a convex arc facing left falling between the item vector and dimension 1.) Because of item 1's proximity to dimension 1, it is primarily measuring θ_1 . Moreover, item 1 is more useful in assessing individuals whose θ_1 locations are in the general neighborhood of -1 to 0 than outside this range. In a proficiency assessment situation, item 1 would be considered to be of average difficulty.

Given that items 1 and 2's vectors are not perpendicular to one another, we know that θ_1 and θ_2 are related to some degree. Each item is measuring a composite of θ_1 and θ_2 , albeit to different degrees, and this relationship is reflected, in part, in ω . In general, item vectors that are clustered together and that are pointing in the same direction are measuring the same relative combination of the F-dimensions. Similarly, a different cluster of item vectors would be measuring a different combination of the F-dimensions. An extreme condition occurs when the item clusters are orthogonal to one another. In this case, the two clusters are measuring completely differently from one another. Implied in the foregoing is that although θ_1 and θ_2 are oriented at 90° to one another, this does not mean that θ_1 and θ_2 are not interrelated. Rather, it is the item vectors' orientation that represents the interrelationship of θ_1 and θ_2 . This is similar to our

use of a scatterplot to graphically depict the interrelation between two variables even though the axes are orthogonal.

In addition to allowing the simultaneous presentation of the locations and discrimination capacity of an item set, a vector plot may be extended to accommodate representing three latent variables. In contrast, with an IRS it is difficult to depict a three-latent-person continua situation (i.e., θ_1 , θ_2 , θ_3). For example, in a three-dimensional figure (e.g., Figure 10.4) each axis would represent one of the three θ_j s (i.e., the X-axis be θ_1 , the Y-axis be θ_2 , and the Z-axis be θ_3). In this case, Equation 10.10 is used to determine the angles of an item vector with respect to a reference plane, such as the $\theta_1 - \theta_2$ plane. With these angles and the item's Δ_j and A_j , we would be able to locate the item in the three-dimensional space. We could also extend this single item representation to an item set.

The Multidimensional Three-Parameter Logistic Model

In the preceding discussion we have focused on the M2PL model. However, the principles outlined can be generalized to other MIRT models. For instance, the M2PL model may be extended to create a multidimensional version of the three-parameter model. This model, the (compensatory) *multidimensional three-parameter logistic* (M3PL) model, is

$$p(x_{ij} = 1 | \underline{\theta}_i, \underline{a}'_j, \gamma_j, \chi_j) = \chi_j + (1 - \chi_j) \frac{e^{\underline{a}'_j \underline{\theta}_i + \delta_j}}{1 + e^{\underline{a}'_j \underline{\theta}_i + \delta_j}} \quad (10.15)$$

where a_{if} and γ_j are defined above, and χ_j is the pseudo-guessing parameter for item j . The χ_j is interpreted analogously to the way it is with the 3PL model.⁷ That is, χ_j is the probability for a response of 1 when an individual is extremely low on all θ s (i.e., $\underline{\theta}_i = -\infty$). Accordingly, the IRS is asymptotic with χ_j . When $\chi_j > 0$, then the corresponding IRS is raised above the “floor” of the graph by an amount equal to χ_j ; in the two-dimensional case the graph’s floor is the $\theta_1 - \theta_2$ plane. Analogous to the situation with the 3PL model, with the M3PL model the inflexion line for the surface corresponds to a probability of $(\chi_j + 1)/2$.

Assumptions of the MIRT Model

As is true with the unidimensional models, MIRT models make a functional form assumption. This assumption states the data follow the function specified by the model. For instance, for Equation 10.3 the functional form states that the probability of a response of 1 increases monotonically when there is an increase in any one or any combination of a person's θ s and that for infinitely low θ s the probability of $\chi_j = 1$ approaches zero.

A second assumption is the conditional independence assumption that is seen with the unidimensional models. It states that for any group of individuals that are charac-

terized by the same values of $\theta_1, \theta_2, \dots, \theta_F$, the conditional distributions of the item responses are all independent of each other (Lord & Novick, 1968). Therefore, whatever relationship exists among the items disappears when one conditions on $\underline{\theta}$.

The third assumption is a dimensionality assumption and states that the observations on the manifest variables are a function of a set of continuous latent person variables. As is the case with the unidimensional models, the proper application of a MIRT model involves dimensionality assessment to determine the number of latent variables to model the response data. For instance, if the true state of nature is that the item responses are a function of three latent variables, then the dimensionality assessment should facilitate correctly specifying that the model has three latent person variables and not, for example, two. As is the case with the unidimensional models, violation of the dimensionality assumption is a matter of degree, and whether the resulting $\underline{\theta}$ s are useful and psychologically meaningful is a validity question.

Estimation of the M2PL Model

Several approaches can be used to estimate the M2PL model's item parameters. One approach uses JMLE (e.g., McKinley & Reckase, 1983b), with the equations that need to be solved presented in McKinley and Reckase (1983a); also see Carlson (1987). A second approach is presented by Bock and Aitkin (1981). Their approach applies MMLE to estimating the item parameters of a multidimensional two-parameter normal ogive model.⁸ MMLE has also been applied to directly estimate the M2PL model (McKinley, 1987; cited in McKinley & Kingston, 1988).

A third approach involves fitting a polynomial that approximates the multidimensional two-parameter normal ogive model. (The extension of the two-parameter normal ogive model to its multidimensional equivalent is discussed in Appendix C.) This model is the multidimensional analog of Equation C.9 (Appendix C) and is the normal ogive version of the model in Equation 10.3. The multidimensional two-parameter normal ogive model states that the proportion of individuals responding 1 on item j , $\pi(x_j = 1)$, is given by

$$\pi(x_j = 1 | \underline{\theta}) = \Phi(\gamma_j + \alpha'_j \underline{\theta}), \quad (10.16)$$

where $\Phi(\bullet)$ is the cumulative normal distribution function (see Fraser & McDonald, 1988; McDonald, 1997, 1999). As explained in Appendix C, the proportion of individuals responding 1 on item j is a function of the area under the unit normal distribution cutoff by the item's threshold, τ_j . By using sample information, we can obtain an initial estimate of τ_j that is subsequently refined to provide an estimate of the intercept (Fraser & McDonald, 2003, 2012; McDonald, 1997). Obtaining estimates of the α_{jj} s involves the observed joint proportion of 1s for items j and v . Specifically, unweighted least squares is used to minimize the squared discrepancies between the observed joint proportion of 1s for items j and v , p_{jv} , and what would be the predicted joint proportion of 1s for items j and v , $\pi_{jv}^{(r)}$

$$F = \sum_{j \neq v} (p_{jv} - \pi_{jv}^{(r)})^2, \quad (10.17)$$

where the term $\pi_{jv}^{(r)}$ is given by an r-term polynomial series with coefficients defined by normalized Hermite–Tchebycheff polynomials (Fraser & McDonald, 1988; McDonald, 1997). (Fraser and McDonald [1988] state that a four-term polynomial is used to determine $\pi_{jv}^{(r)}$.) This approach is implemented in the program NOHARM; see Fraser and McDonald (2003, 2012) for the specifics on the polynomials series. Therefore, when we use NOHARM for parameter estimation, we are, in addition to employing the assumptions mentioned above, assuming that θ is random with an F-variate normal distribution.

Several programs are available for parameter estimation—for example, flexMIRT, mirt, and NOHARM.⁹

Information for the M2PL Model

With the unidimensional models, our items' information for estimating a person's location is a function of their response functions' slopes. In the current context, rather than a response function for describing our item, we have a response surface. Conceptually, this IRS can be viewed as a collection of IRFs that are conditional on a particular direction from the origin. In other words, there is a conditional IRF for 1° , another conditional IRF for 2° , and so on. Applying the item information function from, for example, Equation 2.16 (Chapter 2) to each of these conditional IRFs results in a series of item information functions that are conditional on a direction from the origin. Conceptually, these conditional item information functions collectively form a multidimensional surface. That is, with MIRT models one has a conditional item information *surface* that is called the *multidimensional item information surface*. A graphical depiction of a multidimensional item information surface for the item presented in Figure 10.1 is shown in Figure 10.6.

In the unidimensional case, the relationship between an item's information and the IRF's slope is that, all things being equal, the steeper the IRF's slope, the greater the item information. Conversely, as the slope becomes less steep, one sees a reduction in item information. Similarly, in the multidimensional situation, there is a direct relationship between the slope of the IRS and the amount of item information. However, the multidimensional situation differs from the unidimensional case in that one can calculate different slopes in the θ space conditional on the direction one chooses to traverse the IRS (cf. Figure 10.2).

Recall that through Equation 10.7 we can calculate the value of the slope of the IRS in a direction given by ω_j . Therefore, by extending the unidimensional item information function (Chapter 5: Equation 5.4) to the multidimensional situation and taking into account the directional aspect of determining a slope given by Equation 10.7, we obtain an item's *multidimensional item information*, $I_{j\omega}(\theta)$ (Reckase & McKinley, 1991)

$$I_{j\omega}(\theta) = p_j(1 - p_j)(\sum_{f=1}^F \alpha_{jf} \cos \omega_{jf})^2, \quad (10.18)$$

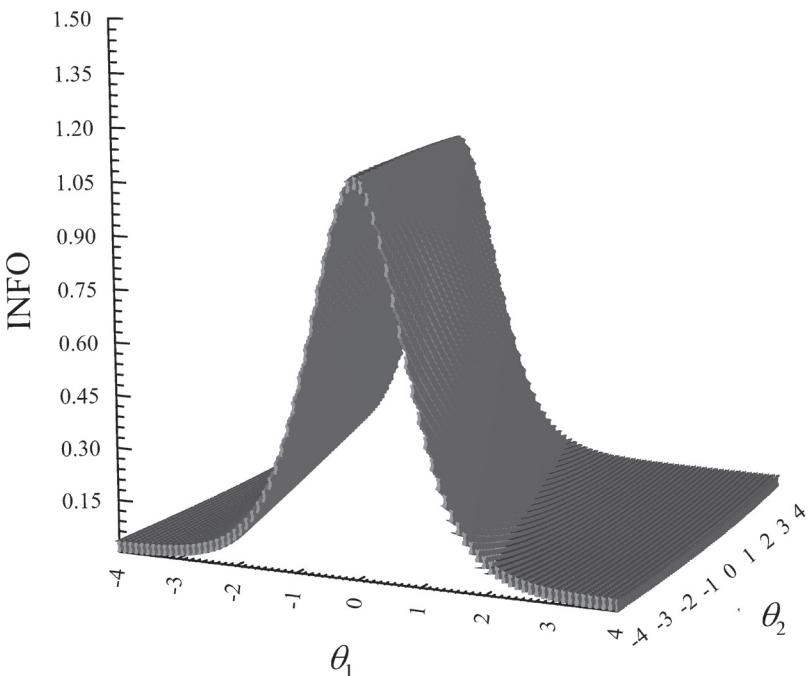


FIGURE 10.6. Multidimensional item information surface for the item in Figure 10.1, $\omega_1 = 14.0362^\circ$.

where p_j is given by the MIRT model and all other terms have been defined above.¹⁰ Equation 10.18 tells us how much information item j can provide for estimation in the direction given by ω_j . This formula shows that item information is calculated with respect to the slope in the direction given by ω_j at the point given by $\boldsymbol{\theta}$. As such, to create the multidimensional item information surface shown in Figure 10.6, we chose its direction to be consistent with the item's vector. Selecting a different direction would potentially result in a different multidimensional item information surface. Therefore, unlike the case with unidimensional models, with MIRT models an item has multiple multidimensional item information surfaces corresponding to different ω_j s. To simultaneously display an item's multiple $I_{j\omega}(\boldsymbol{\theta})$ s, one may use a “clamshell” plot (e.g., Reckase & McKinley, 1991; Yao & Schwarz, 2006). (It should be noted that Equation 10.18 does not take into account the lack of conditional independence that arises from specifying a direction; Ackerman [1994] discusses this issue and presents a solution.)

As is true with the unidimensional models, one may sum the item information to produce a *total multidimensional information* for the instrument, $I_\omega(\boldsymbol{\theta}) = \sum I_{j\omega}(\boldsymbol{\theta})$. However, because the $I_{j\omega}(\boldsymbol{\theta})$ s are defined with respect to a specific direction, the total multidimensional information is also defined with respect to the direction used to determine the multidimensional item information. One may graphically represent the total multidimensional information for the instrument with a *total multidimensional information surface* in the direction ω_j .

Indeterminacy in MIRT

With MIRT there are two sources of indeterminacy. As is the case with the unidimensional models' identification, our first source of indeterminacy in the M2PL model is the indeterminacy of the metric. As discussed in Chapter 3, our metric does not have an intrinsic origin or unit (i.e., the continuum's metric is not absolute, but rather relative). In the MIRT context, this means we have multiple latent person dimensions, each with an indeterminacy of metric issue.

To help conceptualize this metric indeterminacy issue with multiple dimensions, consider Figure 10.4. If we multiply each θ_f by a nonfractional constant (i.e., stretch out the θ_f metric) and divide the corresponding α_{if} s by the same constant, then the plane's orientation in the three-dimensional space does not change. As a consequence, the probabilities do not change and the corresponding IRS is the same as shown in Figure 10.1. (This would also be true if we contract the metric by dividing each θ_f by a nonfractional constant and multiply the corresponding α_{if} s by the same constant.) If we add a constant to γ_j and subtract from each $\alpha_{if}\theta_f$ the ratio of the constant to F, then the plane simply moves up in the three-dimensional space. Because the plane's orientation does not change, the probabilities do not change and the corresponding IRS is the same as shown in Figure 10.1. Similarly, if we subtract a constant from γ_j and add to each $\alpha_{if}\theta_f$ the ratio of the constant to F, then the plane simply moves down but does not change its orientation in the three-dimensional space. Therefore, we can change the values of the $\alpha_{if}\theta$ s, γ_j , and θ without affecting the probability of a response of 1.

As is true with the unidimensional models, how this metric indeterminacy is addressed depends on the estimation approach. For instance, with the JMLE the $\hat{\theta}$ s are rescaled to have a mean of zero and a variance of 1 (i.e., standardized) for each dimension, the $\hat{\alpha}_{if}$ s are multiplied by the standard deviation of the corresponding dimension's $\hat{\theta}$ s, and the $\hat{\gamma}_j$ s are adjusted by $\sum_f \hat{\alpha}_{if} \hat{\theta}_f$. One program for performing MIRT calibrations, NOHARM, addresses the indeterminacy of the metric by ensuring that each component of θ has a mean of zero and a variance of 1 (McDonald, 1997). (NOHARM assumes the multivariate unit normal distribution for θ .) Other programs used for MIRT calibration (e.g., flexMIRT, mirt) use MMLE for estimation. As a result, they address the indeterminacy of metric issue in a fashion analogous to that discussed in Chapter 4 but applied to each dimension.

The second source of indeterminacy is *rotational indeterminacy*. Analogous to factor analysis, the axes' orientation in, say Figure 10.5, are not unique. Rotational indeterminacy, as well as metric indeterminacy, is a reflection of the fact that we have only an internal frame of reference. To understand rotational indeterminacy consider, by way of analogy, a directional compass. Without a magnetic North the compass's needle would freely rotate about its spindle, indicating that "North" is in any direction. Thus, if we use this compass to indicate the direction to, say a mountain, we may find that on its first use the needle points toward the mountain. However, on the compass's second use the needle may point at an angle 25° from where it previously pointed, and with a third use the needle may point 75° from its first orientation. In each instance, if we take into account the magnitude of the angle, we can arrive at the mountain.

In our current context, the axes that represent our latent variables are free to rotate about their origin (i.e., the compass's spindle) because we do not have an external frame of reference to fix their orientation (i.e., we do not have a magnetic North). However, just as in our analogy, if we take into account the angle of orientation, we arrive at the same probability of a response of 1.

As an example, let us use our example item from Figure 10.1 with $\underline{\alpha}'_j = (2.0, 0.5)$ and $\gamma_j = 2.0$. If $\underline{\theta}' = (1.5, -1)$, then our probability of a response of 1 is 0.989. Let us rotate our θ -axes by 25° . When we rotate, our θ -axes our $\underline{\alpha}_j$ and $\underline{\theta}$ are affected. Therefore, after rotation our transformed discrimination parameters become $\alpha_{j1}^* = 2.0239$ and $\alpha_{j2}^* = 1.2984$. Applying the rotation to the θ 's yields $\theta_1^* = 2.7724$ and $\theta_2^* = -2.3962$. Using these transformed discriminations and thetas, $\underline{\alpha}_j^*$ and $\underline{\theta}^*$, the probability of a response of 1 is the same as obtained with the unrotated parameters, 0.989. If we further rotated the axes by 75° , then our $\underline{\alpha}_j^* = (1.006, 2.0613)$ and $\underline{\theta}^{**} = (-1.5636, 1.9719)$. Using these α_j^* 's and this $\underline{\theta}^*$ in Equation 10.3 results in a $p_j = 0.989$. The foregoing shows that we do not have a unique set of $\underline{\alpha}_j'$ and $\underline{\theta}^*$ because of rotational indeterminacy. Stated another way, we can produce an infinite number of θ/α combinations by rotating the axes and obtain the same p_j . Figure 10.7 depicts the rotation of the axes by an angle of Δ as well as the change in item j 's vector before and after the rotation (i.e., $\underline{\alpha}_j^*$).¹¹

To solve the rotational indeterminacy within our compass analogy one could, say, fix the North/South axis to always point North. By fixing this axis, then the East/West axis would also be fixed. Similarly, in the multidimensional context, a strategy for

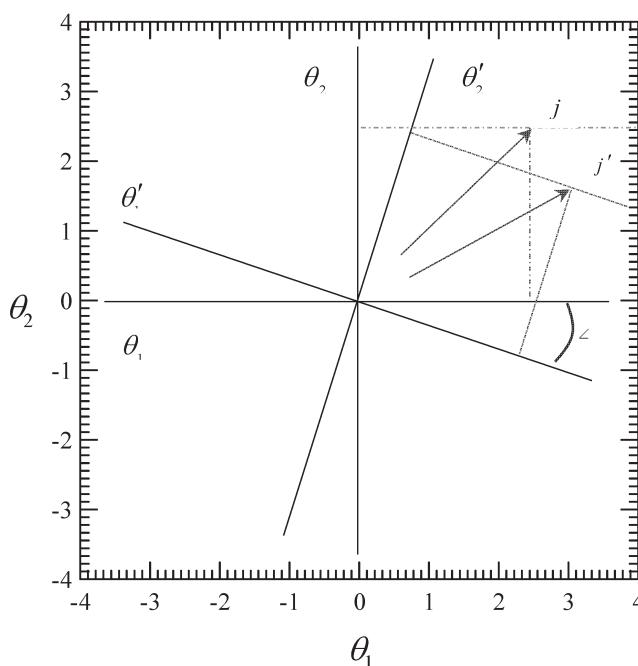


FIGURE 10.7. Axis rotation of two-dimensional structure and its effect on item j .

addressing rotational indeterminacy is to fix the axes by setting one of the items' estimated parameters on one axis to zero. Fixing one axis fixes the other axis.

Although the preceding discussion has presented the rotational indeterminacy identification problem in terms of two dimensions, the issue applies to more than two dimensions. Addressing the problem for more than two dimensions may be accomplished the way it is done for two dimensions. Specifically, for three factors one would restrict one item's estimates on the $F - 1$ factors to be zero, and for a second item one constrains the F th estimate to be 0. With four factors ($F - 1 = 3$) three items would be necessary to address the rotational indeterminacy. Therefore, for one item its loadings on the $F - 1$ factors are restricted to be zero, on a second item one constrains $F - 2$ loadings to be 0, and for a third item one constrains the F th loading to be 0. This is how NOHARM addresses the rotational indeterminacy issue.

Metric Transformation, M2PL Model

The metric transformation equations for the M2PL model are the matrix equivalents of those presented with the two-parameter model (Chapter 5). For example, from Equation 5.6 we have that $\alpha_j^* = \frac{\alpha_j}{\zeta}$. The matrix algebra equivalence for transforming the α_{ij} s (or their estimates) from their initial metric to a target metric is

$$\underline{\alpha}_j^* = (\underline{Z}^{-1})' \underline{\alpha}_j. \quad (10.19)$$

To transform the unidimensional intercepts we use $\gamma_j^* = \gamma_j - \frac{\alpha(\kappa)}{\zeta}$. In terms of matrix algebra, its equivalence is

$$\gamma_j^* = \gamma_j - \underline{\alpha}_j' \underline{Z}^{-1} \underline{\kappa}. \quad (10.20)$$

The matrix equivalent for transforming the person location parameters (or their estimates) of $\theta_i^* = \zeta(\theta_i) + \kappa$ (e.g., Equation 4.17) is

$$\underline{\theta}_i^* = \underline{Z} \underline{\theta}_i + \underline{\kappa}. \quad (10.21)$$

For Equations 10.19–10.21 \underline{Z} is a matrix with dimensions $F \times F$ containing the unit adjustments, its inverse is represented as \underline{Z}^{-1} , and the transpose of the inverse is $(\underline{Z}^{-1})'$. (The inverse of a matrix is the matrix algebra equivalent of division.) The vector $\underline{\kappa}$ is of length F and consists of location adjustments; also see Hirsch (1989).

As an example of applying Equations 10.19–10.21, let $\underline{\alpha}' = (2.0, 0.5)$, $\gamma = 2.0$, and $\underline{\theta} = (1.5, -1.0)$. According to the M2PL model, $p_j = 0.9890$. We transform these parameters by shifting the metric by 1.5 and adjusting the unit by 2. Our metric transformation matrix and vector are

$$\underline{Z} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \text{ and } \underline{\kappa} = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}.$$

The inverse of $\underline{\mathbf{Z}}$ is

$$\underline{\mathbf{Z}}^{-1} = \begin{bmatrix} 0.6667 & -0.3333 \\ -0.3333 & 0.6667 \end{bmatrix}.$$

In this example the transpose of $\underline{\mathbf{Z}}^{-1}$ is the same as $\underline{\mathbf{Z}}^{-1}$ (i.e., $(\underline{\mathbf{Z}}^{-1})' = \underline{\mathbf{Z}}^{-1}$). Applying Equation 10.19 to transform the α_{jfs} , we obtain

$$\underline{\mathbf{a}}_j^* = (\underline{\mathbf{Z}}^{-1})' \underline{\mathbf{a}}_j = \begin{bmatrix} 0.6667 & -0.3333 \\ -0.3333 & 0.6667 \end{bmatrix} \begin{bmatrix} 2.0 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1.1667 \\ -0.3333 \end{bmatrix}.$$

and for the intercept term we have

$$\gamma_j^* = \gamma_j - \underline{\mathbf{a}}_j' \underline{\mathbf{Z}}^{-1} \underline{\mathbf{k}} = 2.0 - [2.0 \quad 0.5] \begin{bmatrix} 0.6667 & -0.3333 \\ -0.3333 & 0.6667 \end{bmatrix} \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix} = 0.75.$$

Transforming our θ s produces

$$\underline{\boldsymbol{\theta}}_i^* = \underline{\mathbf{Z}} \underline{\boldsymbol{\theta}}_i + \underline{\mathbf{k}} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1.5 \\ -1.0 \end{bmatrix} + \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 1.0 \end{bmatrix}.$$

As would be expected, given the invariance of our parameters our use of α^* , γ^* , and $\boldsymbol{\theta}^*$ with the M2PL model yields $p_j = 0.9890$.

With the 2PL model the total characteristic curve could be used to transform our θ (or its estimate) to an expected trait score. For the M2PL model the total characteristic curve becomes the *total characteristic surface*. This surface is defined as

$$\mathcal{ET} = \frac{\sum p_j}{L}, \quad (10.22)$$

where p_j is a MIRT model (Reckase, 1997a; Reckase, Ackerman, & Carlson, 1988).

Example: Calibration of Interpersonal Engagement Instrument, M2PL Model, `sirt.noharm`

Assume that we have a 10-item instrument designed to measure interpersonal engagement behavior and we administer it to 1000 individuals. This instrument uses a true/false response format. Item responses are assumed to be a function of an individual's locations on the "love" and "dominance" dimensions. Accordingly, we believe two dimensions underlie the response data. However, following practice we assess the fit of models with a varying number of latent variables to the data.

We use NOHARM to fit a one-, two-, and three-dimensional two-parameter model to the data. We use the NOHARM function from the R package `sirt` here; Appendix G, "Standalone NOHARM Calibration of Interpersonal Engagement Instrument, M2PL

Model,” shows the analysis using the stand-alone NOHARM program. Our R session is shown in Table 10.1.

As described in Chapter 3, our session begins with loading the `sirt` package (`library(sirt)`) and reading our data. To verify that the data are read correctly, we display the first and last five cases of interpersonal data using the `head` and `tail` functions. In our call to `noharm.sirt(. . .)`, we provide our data (`intprnsndat`) as the first argument, the fitting of a unidimensional model (`dimensions = 1`), that the IRF lower asymptotes for each item is 0 (`lower = 0`; i.e., we are assuming the response data are not influenced by guessing), to use the `optim` optimizer, and to calculate the reliability (`reliability = TRUE`).

We obtain our results using the `summary` function (`summary(noharm1d)`). The “Information about optimization” section shows that the correct number of cases and items were used, that the dimensionality is what we want, and that we obtained a converged solution (`Converged = TRUE` and `Number of iterations = 13`); see the “Information about Optimization” section.

Table 10.2 summarizes our NOHARM fit information in its top panel. Our one-dimensional solution shows a GFI of 0.99937, RMSEA = 0.036, and a SRMSR = 0.02918 (obtained by `modelfit.sirt(noharm1d)`). Given the guidelines, these values all indicate model-data fit. This is corroborated by a RMSR (0.00635) that is substantially less than the target value of 0.12649. Examination of our residual matrix (`noharm1d$residuals`) shows the residuals range from -0.01029 to 0.01221. Thus, our reproduced values are very close to what was observed. In light of the residuals and according to our indices, there does not appear to be sufficient evidence to reject a unidimensional solution. However, we note that our significant G&D Chi Square statistic (i.e., at the 1% significance level) contradicts this interpretation.

Our two-dimensional (`noharm2d = noharm.sirt(. . . , dimensions = 2, . . .)`) and three-dimensional (`noharm3d = noharm.sirt(. . . , dimensions = 3, . . .)`) models all exhibit model-data fit according to GFI, RMSR, RMSEA, and SRMSR. The G&D Chi Square statistics are not significant and are significantly less than the unidimensional model’s G&D Chi Square of 79.645. Our residuals show a range that is less than those from the unidimensional model. Specifically, for the two-dimensional model our minimum residual is -0.00775 and our maximum residual is 0.00981, whereas for the three-dimensional model they are `min = -0.00510` and `max = 0.00361`. That is, our predicted covariances are closer to the observed covariances using the two-dimensional model than with the one-dimensional model, whereas with the three-dimensional model there is better correspondence than with the one- and two-dimensional models.

An argument can be made that the additional complexity of the three-dimensional model is not necessary given the unidimensional and two-dimensional models’ fit vis-à-vis their GFI, RMSR, RMSEA, and SRMSR values. Moreover, in practice dimensional interpretation would help determine whether a three-dimensional model is appropriate. In terms of the G&D Chi Square, the two-dimensional model’s value is the simplest multidimensional model that is nonsignificant. Moreover, our theoretical understand-

TABLE 10.1. sirt.noharm Session for MIRT Analysis

```

> library(sirt)
- sirt 3.4-64 (2019-05-03 18:33:11)

> intprnsndat=read.table("intrprsnl.dat",col.names=c(paste0("i",1:10)))

> head(intprnsndat,5)
  i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
1 1 1 1 0 1 0 0 0 0 0
2 1 1 0 1 1 0 0 0 0 0
3 1 1 1 0 0 0 0 0 0 1
4 1 1 0 1 0 0 1 0 0 0
5 1 1 0 1 0 0 0 1 0 0

> tail(intprnsndat,5)
  i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
996 1 1 1 0 0 0 0 0 0 0
997 0 0 1 0 0 0 0 0 0 0
998 1 0 0 1 1 0 1 1 0 0
999 1 1 1 1 1 1 1 1 1 1
1000 1 0 0 1 0 0 0 0 0 0

> # One dimensional solution =====
> noharm1d=noharm.sirt(intprnsndat, dimensions=1,lower=0, optimizer="optim",
  reliability=TRUE)
> summary(noharm1d)
-----
sirt 3.4-64 (2019-05-03 18:33:11)
R version 3.6.0 (2019-04-26) i386, mingw32

Call:
noharm.sirt(dat = intprnsndat, dimensions = 1, lower = 0, optimizer = "optim",
  reliability = TRUE)
:
--- Information about optimization ---
Optimizer = optim
Converged = TRUE
Optimization Function Value = 0.001814
Number of iterations = 13
:
Number of Observations: 1000
Number of Items : 10
Number of Dimensions : 1
Tanaka Index : 0.99937
RMSR : 0.00635
                                         ← The GFI
                                         ← The RMSR

Number of Used Item Pairs : 45
Number of Estimated Parameters : 20
  # Thresholds : 10
  # Loadings : 10
  # Variances/Covariances : 0
  # Residual Correlations : 0

Chi Square Statistic of Gessaroli & De Champlain (1996)
Chi2 : 79.645
Degrees of Freedom (df) : 35
p(Chi2,df) : 0
Chi2 / df : 2.276
RMSEA : 0.036
                                         ←
                                         ← The RMSEA

```

(continued)

TABLE 10.1. (*continued*)

```

Green-Yang Reliability Omega Total : 0.721

Factor Covariance Matrix
  F1
F1  1

Factor Correlation Matrix
  F1
F1  1

Item Parameters - Latent Trait Model (THETA) Parametrization
  Loadings, Constants, Asymptotes and Descriptives
    F1 final.constant lower upper item.variance   N      p
  i1  1.161          0.821     0     1        2.348 1000 0.704
  i2  0.638          0.170     0     1        1.407 1000 0.557
  :
  i10 0.674         -1.019    0     1        1.455 1000 0.199

Item Parameters - Common Factor (DELTA) Parametrization
  Loadings, Thresholds, Uniquenesses and Asymptotes
    F1 threshold lower upper uniqueness
  i1  0.758         -0.536     0     1        0.426
  i2  0.538         -0.143     0     1        0.711
  :
  i10 0.559         0.845     0     1        0.687

--- Parameter table ---
  mat row col index fixed   est lower
  1   F   1   1       1      0 1.161  -Inf
  2   F   2   1       2      0 0.638  -Inf
  :
  10  F   10  1       10     0 0.674  -Inf
  11  P   1   1       NA     1 1.000  NA

> summary(modelfit.sirt(noharm1d))
Test of Global Model Fit
  type   value p
  1 max(X2) 3.39910 1
  2 abs(fcor) 0.06347 1

Fit Statistics
  est
MADcor      0.02483
SRMSR       0.02918
100*MADRESIDCOV 0.52950 ←
MADQ3       0.07648
MADAQ3      0.03486

> noharm1d$residuals
      i1        i2        i3        i4        i10
  i1  0.000000000  0.0103335850  0.0089999945 -0.0043103973 -0.0022834260
  i2  0.010333585  0.0000000000  0.0110458777 -0.0030084260  0.0004842495
  i3  0.008999995  0.0110458777  0.0000000000  0.0041855311 -0.0007957037
  i4 -0.004310397 -0.0030084260  0.0041855311  0.0000000000  0.0049723678
  i5  0.008743130  0.0008744472  0.0015214252 -0.0009122786 -0.0049878064
  i6 -0.003905716 -0.0045855861 -0.0080897488 -0.0002463333 -0.0024090396
  i7 -0.006868218 -0.0092907869 -0.0067341806 -0.0060499479 -0.0025626230
  i8 -0.009484910 -0.0068839553 -0.0102910127  0.0072474034  0.0107933170
  i9 -0.007409471 -0.0042717368 -0.0062452997  0.0021891440 -0.0028917707
  i10 -0.002283426  0.0004842495 -0.0007957037  0.0049723678  0.0000000000

```

(continued)

TABLE 10.1. (continued)

```

> # Two dimensional solution =====
> noharm2d=noharm.sirt(intprnsndat,dimensions=2,lower=0,optimizer="optim",
+ reliability=TRUE)
> summary(noharm2d)
-----
:
Call:
noharm.sirt(dat = intprnsndat, dimensions = 2, lower = 0, optimizer = "optim",
reliability = TRUE)
:
--- Information about optimization ---
Optimizer = optim
Converged = TRUE
Optimization Function Value = 0.000506
Number of iterations = 30
:
Number of Dimensions : 2
Tanaka Index : 0.99982
RMSR : 0.00335

Number of Used Item Pairs : 45
Number of Estimated Parameters : 29
# Thresholds : 10
# Loadings : 19
# Variances/Covariances : 0
# Residual Correlations : 0

Chi Square Statistic of Gessaroli & De Champlain (1996)
Chi2 : 23.469
Degrees of Freedom (df) : 26
p(Chi2,df) : 0.606
Chi2 / df : 0.903
RMSEA : 0

Green-Yang Reliability Omega Total : 0.73

Factor Covariance Matrix
F1   F2
F1 1.000 0.759
F2 0.759 1.000

Item Parameters - Promax Rotated Parameters (THETA)
Loadings, Constants, Asymptotes and Descriptives
      F1      F2 final.constant lower upper    N     p
i1  0.000  0.873          0.912    0    1 1000 0.704
i2 -0.033  0.657          0.177    0    1 1000 0.557
:
i10 0.397  0.241         -1.018    0    1 1000 0.199

Item Parameters - Promax Rotated Parameters (DELTA)
Loadings, Constants, Asymptotes and Descriptives
      F1      F2 thresh lower upper    N     p
i1  0.000  0.809 -0.536     0    1 1000 0.704
i2 -0.030  0.609 -0.143     0    1 1000 0.557
:
i10 0.368  0.224  0.845     0    1 1000 0.199

--- Parameter table ---
  mat row col index fixed   est lower
  1   F   1   1     1     0 0.046  -Inf
  2   F   3   1     2     0 0.009  -Inf

```

(continued)

TABLE 10.1. (*continued*)

```

:
20   P   1   1     NA      1 1.000    NA
21   P   2   2     NA      1 1.000    NA

> summary(modelfit.sirt(noharm2d))
  Test of Global Model Fit
    type   value      p
  1  max(X2) 8.85781 0.13133
  2 abs(fcor) 0.08302 0.19704

  Fit Statistics
                est
MADcor        0.03937
SRMSR         0.04514
100*MADRESIDCOV 0.86517
MADQ3          0.09047
MADAQ3         0.03361

> noharm2d$residuals
      i1           i2           i3           i4           i10
i1  0.000000000 2.560651e-04 -0.0018017183 -0.0031371202 -0.0009643926
i2  0.0002560651 0.000000e+00  0.0015934895 -0.0023963296  0.0015859411
i3 -0.0018017183 1.593489e-03 0.0000000000  0.0050236615  0.0004375241
i4 -0.0031371202 -2.396330e-03 0.0050236615 0.0000000000  0.0050166016
i5  0.0051291078 -2.661629e-03 -0.0020472648 -0.0005235451 -0.0043177621
i6  0.0021356074  1.093592e-03 -0.0021361198 -0.0007198331 -0.0029183974
i7  0.0013889465 -1.771525e-03  0.0011954403 -0.0077508401 -0.0042212134
i8 -0.0020799956 -4.472041e-05 -0.0031128062  0.0063350209  0.0098098323
i9 -0.0002925729  2.735193e-03  0.0009730578  0.0017753801 -0.0035035428
i10 -0.0009643926 1.585941e-03  0.0004375241  0.0050166016  0.0000000000

> # Three dimensional solution =====
> noharm3d=noharm.sirt(intprnsndat,dimensions=3,lower=0,optimizer="optim",
  reliability=TRUE)
> summary(noharm3d)
-----
:
Call:
noharm.sirt(dat = intprnsndat, dimensions = 3, lower = 0, optimizer = "optim",
  reliability = TRUE)
:
--- Information about optimization ---
Optimizer = optim
Converged = TRUE
Optimization Function Value = 0.00017
Number of iterations = 81
:
Number of Dimensions : 3
Tanaka Index          : 0.99994
RMSR                 : 0.00194

Number of Used Item Pairs : 45
Number of Estimated Parameters : 37
  # Thresholds          : 10
  # Loadings            : 27
  # Variances/Covariances : 0
  # Residual Correlations : 0

Chi Square Statistic of Gessaroli & De Champlain (1996)
Chi2                  : 7.265
Degrees of Freedom (df) : 18

```

(continued)

TABLE 10.1. (continued)

```

p(Chi2,df) : 0.988
Chi2 / df   : 0.404
RMSEA      : 0

Green-Yang Reliability Omega Total : 0.737

Factor Covariance Matrix
  F1    F2    F3
F1 1.000 0.738 0.745
F2 0.738 1.000 0.635
F3 0.745 0.635 1.000

Item Parameters - Promax Rotated Parameters (THETA)
 Loadings, Constants, Asymptotes and Descriptives
  F1    F2    F3 final.constant lower upper   N     p
i1 -0.087 0.088 0.910      0.940      0     1 1000 0.704
i2  0.088 -0.053 0.606      0.176      0     1 1000 0.557
:
i10 0.727 -0.110 0.054     -1.087      0     1 1000 0.199

Item Parameters - Promax Rotated Parameters (DELTA)
 Loadings, Constants, Asymptotes and Descriptives
  F1    F2    F3 thresh lower upper   N     p
i1 -0.079 0.080 0.828 -0.536      0     1 1000 0.704
i2  0.080 -0.048 0.551 -0.143      0     1 1000 0.557
:
i10 0.662 -0.100 0.049  0.845      0     1 1000 0.199

--- Parameter table ---
  mat row col index fixed   est lower
1   F   1   1       1      0 -0.096 -Inf
2   F   4   1       2      0  0.359 -Inf
:
28  P   1   1       NA     1  1.000  NA
29  P   2   2       NA     1  1.000  NA
30  P   3   3       NA     1  1.000  NA

> summary(modelfit.sirt(noharm3d))
Test of Global Model Fit
  type   value     p
1 max(X2) 21.03504 0.00020
2 abs(fcor) 0.12391 0.00205

Fit Statistics
  est
MADcor        0.03778
SRMSR         0.04777
100*MADRESIDCOV 0.91705
MADQ3          0.05135
MADAQ3         0.02907

> noharm3d$residuals
      i1          i2          i3          i4          i10
i1 0.000000000 4.243092e-04 -2.077344e-03 -0.001251169 1.151271e-03
i2 0.0004243092 0.000000e+00 2.103994e-03 -0.003526242 1.206217e-03
i3 -0.0020773438 2.103994e-03 0.000000e+00 0.003608362 -7.477768e-05
i4 -0.0012511692 -3.526242e-03 3.608362e-03 0.000000000 -1.523572e-03
i5  0.0035435456 -2.275580e-03 -1.741828e-03 0.001666091 -1.882844e-03
i6  0.0010617365 8.889959e-04 -2.063248e-03 0.001325194 -7.355955e-04
i7 -0.0013669431 -1.199955e-03 2.409118e-03 -0.002230139 1.475325e-03

```

(continued)

TABLE 10.1. (continued)

```

i8  0.0010326738 -4.521528e-05 -2.941417e-03 -0.000363561    1.870377e-03
i9  -0.0009202849  2.377016e-03  9.173193e-04  0.003017630   -2.348904e-03
i10 0.0011512709  1.206217e-03 -7.477768e-05 -0.001523572   0.000000e+00

> # return to two-dimensional solution =====
> noharm2d$final.constants                                # our  $\hat{\gamma}_i$ s
      i1          i2          i3          i4          i5
  0.9105733  0.1770683  0.3817659  0.3244944  0.0308340
      i6          i7          i8          i9          i10
 -0.8279098 -0.5197396 -0.6269490 -0.9919529 -1.0179514

> noharm2d$loadings.theta                               # unrotated
      F1          F2
  i1  0.046874565  1.3727616
  i2  0.000000000  0.7248373
  i3  0.009043244  0.8577830
  i4  0.273138953  0.5716664
  i5  0.095734891  0.4909676
  i6  0.586488493  0.7195368
  i7  0.631218123  0.6290609
  i8  0.523282985  0.5208607
  i9  0.606686416  0.5710550
 i10 0.308167815  0.5963183

> noharm2d$thresholds
      i1          i2          i3          i4          i5
 -0.5359400 -0.1433674 -0.2897598 -0.2741101 -0.0275764
      i6          i7          i8          i9          i10
  0.6067754  0.3880217  0.5043720  0.7621005  0.8451985

> noharm2d$uniquenesses
      i1          i2          i3          i4          i5
  0.3464197  0.6555704  0.5760788  0.7135684  0.7998628
      i6          i7          i8          i9          i10
  0.5371429  0.5573658  0.6471985  0.5902588  0.6893875

> varimax(noharm2d$loadings.theta)                      # external function to rotate loadings
  $loadings
  Loadings:
      F1          F2
  i1  0.563  1.256
  i2  0.274  0.671
  i3  0.333  0.790
  i4  0.469  0.426
  i5  0.274  0.418
  i6  0.815  0.444
  i7  0.822  0.344
  i8  0.682  0.284
  i9  0.777  0.300
 i10 0.511  0.435

      F1          F2
SS loadings     3.467  3.685
Proportion Var  0.347  0.368
Cumulative Var 0.347  0.715

$rotmat
      [,1]      [,2]
[1,] 0.9257919 -0.3780336
[2,] 0.3780336  0.9257919

> noharm2d$loadings

```

TABLE 10.2. Comparative Fit Information from NOHARM, mirt, and flexMIRT

sirt.noharm										
	GFI	RMSR	RMSEA	SRMSR	χ^2	p	Nparms	df	$\chi^2_{difference}$	p
1D	0.99937	0.00635	0.036	0.02918	79.645	0.000025	20			
2D	0.99982	0.00335	0	0.04514	23.469	0.606300	29	9	56.176	0.000
3D	0.99994	0.00194	0	0.04777	7.265	0.987695	37	8	16.204	0.040
mirt										
	AIC	BIC	lnL		Nparms	-2lnL		df	ΔG^2	p
1D	11457.74	11555.89	-5708.870		20	11417.74				
2D	11444.89	11597.03	-5691.444		31	11382.89		11	34.852	0.000
3D	11459.81	11670.85	-5686.907		43	11373.81		12	9.074	0.697
	M2	df	p	RMSEA	RMSEA_2.5	RMSEA_97.5	SRMSR	TLI	CFI	
1D	52.592	35	0.028	0.022	0.008	0.036	0.029	0.987	0.990	
2D	17.587	24	0.823	0.000	0.000	0.020	0.017	1.00 ^a	1.000	
3D	4.782	12	0.965	0.000	0.000	0.009	0.013	1.00 ^b	1.000	
flexMIRT										
	AIC	BIC	Nparms		-2lnL		df	ΔG^2	p	RMSEA
1D	11457.30	11555.45	20		11417.30					0.04
2D	11437.72	11582.05	29		11381.72		9	35.58	0.000	0.03
3D	11444.39	11625.98	37		11370.39		8	11.33	0.184	0.04

^aTLI=1.007 set to 1
^bTLI=1.015 set to 1

ing is that responses to our instrument should reflect two dimensions, “love” and “dominance.” Hence, we accept the two-dimensional representation of our data.

Although our M2PL model’s intercepts may be found in different places (e.g., the Promax Rotated Parameters (THETA) section final.constant column) we simply request them (noharm2d\$final.constants). The items’ intercept (constant) estimates, $\hat{\gamma}_j$ s, are $\hat{\gamma}_1 = 0.9106$, $\hat{\gamma}_2 = 0.1771$, ..., $\hat{\gamma}_{10} = -1.0179$. We can obtain our items’ discrimination parameter estimates, $\hat{\alpha}_{ij}$ s, in a similar fashion (noharm2d\$loadings.theta). These values are not rotated and can potentially be negative. Thus, we use the varimax function to rotate our solution (varimax(noharm2d\$loadings.theta)). (We use an orthogonal rotation because our indices (e.g., Δ_j) are based on orthogonal axes, not because we necessarily believe the dimensions are independent. As mentioned

above, the interrelationship of θ_1 and θ_2 is not reflected in the orientation of the axes with respect to one another, but in the item vectors' orientation.)

Table 10.3 presents our estimates. The item discrimination parameter estimates for the first dimension are $\hat{\alpha}_{1,1} = 0.563$, $\hat{\alpha}_{2,1} = 0.274, \dots, \hat{\alpha}_{10,1} = 0.511$ (i.e., column 1). For the second dimension the estimates of the discrimination parameters are $\hat{\alpha}_{1,2} = 1.256$, $\hat{\alpha}_{2,2} = 0.671, \dots, \hat{\alpha}_{10,2} = 0.435$. Applying Equations 10.12, 10.5, and 10.10 yields \hat{A}_j , $\hat{\Delta}_j$, and ω_{j1}^o , respectively. As can be seen, item 1 is the most discriminating item ($\hat{A}_1 = 1.376$) and is slightly below average in difficulty to endorse ($\hat{\Delta}_1 = -0.662$). In contrast, item 10 is our most difficult item to endorse with a $\hat{\Delta}_{10} = 1.517$, whereas item 5 is our least discriminating item with $\hat{A}_5 = 0.5$ and is about average in difficulty to endorse ($\hat{\Delta}_5 = -0.062$). Because these estimates are on the normal metric, we multiply them by $D = 1.702$ to place them on the logistic metric of the M2PL model.

Figure 10.8 contains the vector plot summarizing our instrument. As can be seen, we have a cluster of items (items 4 and 10) in quadrant I that are measuring both dimensions to about the same degree for individuals who are above average in their interpersonal engagement. Most of the remaining items either primarily assess θ_1 (items 6, 7, 8, and 9) or primarily assess θ_2 (items 1, 2, 3, and 5). Because a vector's length reflects its discrimination capacity, we see item 1's vector is the longest ($\hat{A}_1 = 1.376$) and item 5's is the shortest ($\hat{A}_5 = 0.5$). Moreover, because item 10 is the most difficult item to endorse, its starting point is to the right of all the other item vectors. We can attach meaning or labels to θ_1 and θ_2 by interpreting the rotated loadings.

As is clear from the output, NOHARM does not provide standard errors for its estimates. If standard errors are desired, they can be obtained by using the standard errors

TABLE 10.3. Summary Statistics for the Interpersonal Engagement Behavior Instrument

	normal metric		\hat{r}_i	\hat{A}_i	$\hat{\Delta}_i$	ω_{j2}^o	logistic metric	
	$\hat{\alpha}_{j1}$	$\hat{\alpha}_{j2}$					$\hat{\alpha}_{j1}$	$\hat{\alpha}_{j2}$
1	0.563	1.256	0.911	1.376	-0.662	24.1	0.958	2.138
2	0.274	0.671	0.177	0.725	-0.244	22.2	0.466	1.142
3	0.333	0.790	0.382	0.857	-0.445	22.9	0.567	1.345
4	0.469	0.426	0.324	0.634	-0.512	47.8	0.798	0.725
5	0.274	0.418	0.031	0.500	-0.062	33.2	0.466	0.711
6	0.815	0.444	-0.828	0.928	0.892	61.4	1.387	0.756
7	0.822	0.344	-0.520	0.891	0.583	67.3	1.399	0.585
8	0.682	0.284	-0.627	0.739	0.849	67.4	1.161	0.483
9	0.777	0.300	-0.992	0.833	1.191	68.9	1.322	0.511
10	0.511	0.435	-1.018	0.671	1.517	49.6	0.870	0.740

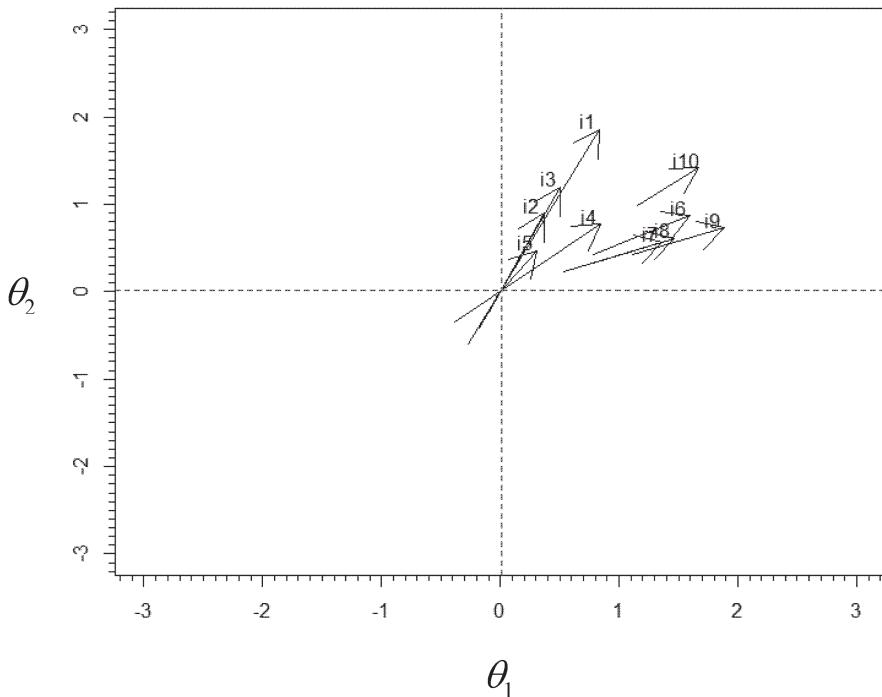


FIGURE 10.8. Vector plot for the interpersonal engagement instrument.

formulas provided by Maydeu-Olivares (2001) or by using the bootstrap methodology (Efron & Tibshirani, 1993).

Obtaining Person Location Estimates

Once we have satisfactory model–data fit, we can estimate the person locations by using, for example, EAP (Bock & Aitkin, 1981). For instance, assuming a two-dimensional latent space, the EAP estimate of person i 's location on dimension 1 is

$$\hat{\theta}_{il} = \sum_{r_1=1}^{R_1} X_{r_1,1} \left[\sum_{r_2=1}^{R_2} L_i(X_{r_1,1}, X_{r_2,2}) A(X_{r_2,2}) \right] \frac{A(X_{r_1,1})}{\tilde{p}_i}, \quad (10.23)$$

where the number of quadrature points on dimensions 1 and 2 are R_1 and R_2 , respectively, $(X_{r_1,1})$ is the r th quadrature point on dimension 1 with corresponding weight $A(X_{r_1,1})$, $X_{r_2,2}$ is the r th quadrature point on dimension 2 with corresponding weight $A(X_{r_2,2})$, L_i is the likelihood function for person i based on the model and evaluated at the $X_{r_1,1}$ and $X_{r_2,2}$ points, and \tilde{p}_i is the unconditional probability of person i 's response vector and is calculated using a two-dimensional Hermite–Gauss quadrature

$$\sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} L_i(X_{r_1,1}, X_{r_2,2}) A(X_{r_1,1}) A(X_{r_2,2}). \quad (10.24)$$

The estimate's standard error is given by

$$PSD(\hat{\theta}_{i1}) = \sqrt{\frac{\sum_{r_1=1}^{R_1} (X_{r,1} - \hat{\theta}_{i1})^2 \left(\sum_{r_2=1}^{R_2} L_i(X_{r,1}, X_{r,2}) A(X_{r,2}) \right) (A(X_{r,1}))}{\sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} L_i(X_{r,1}, X_{r,2}) A(X_{r,1}) A(X_{r,2})}}. \quad (10.25)$$

By appropriate substitution for the points and weights, we have person i 's estimated location on dimension 2

$$\hat{\theta}_{i2} = \sum_{r_2=1}^{R_2} X_{r,2} \left[\sum_{r_1=1}^{R_1} L_i(X_{r,1}, X_{r,2}) A(X_{r,1}) \right] \frac{A(X_{r,2})}{\tilde{p}_i}. \quad (10.26)$$

Example: Calibration of Interpersonal Engagement Instrument, M2PL Model, mirt

Table 10.4 shows our R session using `mirt` to perform a reanalysis of our interpersonal engagement data. As with `sirt.noharm`, we fit one-, two-, and three-dimensional compensatory M2PL models. For our one-dimensional model we fit the 2PL model as done in Chapter 5. For our two- and three-dimensional M2PL models we first need to specify the model's loading structure using `mirt.model`. For example, given our compensatory perspective, we indicate that the ten items load on each factor, F1 and F2 ("F1 = 1-10" and "F2 = 1-10"). Furthermore, we estimate the covariance between the two factors ($\text{COV} = F1 * F2'$). In our call to `mirt`, we pass our model specification (`ModelSpec2D`) and set `itemtype` to be "2PL." Similarly, for our three-dimensional model, we specify that the items load on each factor ($F1 = 1-10, F2 = 1-10, F3 = 1-10$) and to estimate the standardized covariance among the factors ($\text{COV} = F1 * F2 * F3'$).

Table 10.2's `mirt` section contains the model-level fit information. The RMSEA 95% CIs are comparatively narrow for each of our dimensional models; we can expect the true value to be within this range 95% of the time. Because a RMSEA greater than 0.08 is indicative of "mediocre/poor" fit, it is desirable that the RMSEA value and its CI's upper bound are less than 0.08. All of the models' 95% CIs indicate that we are in the "close fit" to "good fit" neighborhood. Because the models' SRMRs are less than the "close to" cutoff of 0.08, these also reflect good fit for each model. Moreover, our TLI and CFI values are each greater than the 0.95 guideline. Based on these indices, our one-, two-, and three-dimensional models fit the data, albeit to varying degrees.

With respect to our information criteria, BIC indicates that the one-dimensional model is a better fit than either the two- or three-dimensional models. In contrast, our AIC, M_2 , and G^2 indicate that the two-dimensional model is a better fit than the one-dimensional model. Moreover, the additional complexity of the three-dimensional model over the two-dimensional model is not warranted on the basis of the AIC, M_2 , and G^2 . For the reasons indicated above with `sirt.noharm` our preferred model is the two-dimensional M2PL model.¹² At the item-level the $S - X^2$ fit statistics are, except for

TABLE 10.4. mirt Session for MIRT Analysis

```

> # load mirt
> # One dimensional solution =====
> print(TwoPL = mirt(intprnsndat,1,'2PL', method = 'MHRM',SE=T)))
  Stage 3 = 48, LL = -6475.4, AR(5.00) = [0.28], gam = 0.0099, Max-Change = 0.0009
  Calculating information matrix...
  Calculating log-likelihood...
  Call:
  mirt(data = intprnsndat, model = 1, itemtype = "2PL", SE = T,
    method = "MHRM")
  Full-information item factor analysis with 1 factor(s).
  Converged within 0.001 tolerance after 48 MHRM iterations.
  mirt version: 1.30
  M-step optimizer: NR1
  Latent density type: Gaussian
  Information matrix estimated with method: MHRM
  Condition number of information matrix = 12.34404
  Second-order test: model is a possible local maximum
  Log-likelihood = -5708.87, SE = 0.021
  Estimated parameters: 20
  AIC = 11457.74; AICc = 11458.6
  BIC = 11555.89; SABIC = 11492.37
  G2 (1003) = 738.46, p = 1
  RMSEA = 0, CFI = NaN, TLI = NaN

> M2(TwoPL,CI=0.95)
      M2 df      p      RMSEA RMSEA_2.5 RMSEA_97.5      SRMSR       TLI       CFI
  stats 52.59169 35 0.02843686 0.0224304          0 0.03633177 0.02856338 0.9874579 0.990245

> coef(TwoPL,simplify=TRUE)
  $items
    a1      d g u
  i1 2.013  1.413 0 1
  i2 1.052  0.276 0 1
  i3 1.263  0.602 0 1
  i4 1.060  0.535 0 1
  i5 0.808  0.045 0 1
  i6 1.479 -1.379 0 1
  i7 1.275 -0.826 0 1
  i8 1.087 -1.013 0 1
  i9 1.221 -1.602 0 1
  i10 1.165 -1.750 0 1

  $means
  F1
  0

  $cov
    F1
  F1  1

> # Two dimensional solution =====
> ModelSpec2D=mirt.model('F1 = 1-10
+ F2 = 1-10
+ COV = F1*F2')

> print((M2PL2D = mirt(intprnsndat,ModelSpec2D,'2PL', method = 'MHRM',SE=T)))
  Stage 3 = 69, LL = -7713.5, AR(1.80) = [0.30], gam = 0.0075, Max-Change = 0.0009
  Calculating information matrix...

```

(continued)

TABLE 10.4. (*continued*)

```

Calculating log-likelihood...

Call:
mirt(data = intprnsndat, model = ModelSpec2D, itemtype = "2PL",
      SE = T, method = "MHRM")

Full-information item factor analysis with 2 factor(s).
Converged within 0.001 tolerance after 69 MHRM iterations.
mirt version: 1.30
M-step optimizer: NR1
Latent density type: Gaussian

Information matrix estimated with method: MHRM
Condition number of information matrix = 248.0876
Second-order test: model is not a maximum, or the information matrix is too inaccurate

Log-likelihood = -5691.444, SE = 0.021
Estimated parameters: 31
AIC = 11444.89; AICc = 11446.94
BIC = 11597.03; SABIC = 11498.57
G2 (992) = 703.51, p = 1
RMSEA = 0, CFI = NaN, TLI = NaN

> M2 (M2PL2D,CI=0.95)
      M2 df          p RMSEA RMSEA_2.5 RMSEA_97.5      SRMSR       TLI   CFI
stats 17.58668 24 0.822586      0      0.01960563 0.01700509 1.006668   1

> # 1D vs 2D model-level fit comparison -----
> anova(TwoPL,M2PL2D)
  Model 1: mirt(data = intprnsndat, model = 1, itemtype = "2PL", SE = T,
                 method = "MHRM")
  Model 2: mirt(data = intprnsndat, model = ModelSpec2D, itemtype = "2PL",
                 SE = T, method = "MHRM")

      AIC      AICc     SABIC      HQ      BIC      logLik      X2    df      p
  1 11457.74 11458.60 11492.37 11495.05 11555.89 -5708.870      NaN  NaN  NaN
  2 11444.89 11446.94 11498.57 11502.71 11597.03 -5691.444 34.85  11    0

> itemfit(M2PL2D,fit_stats="S_X2")
  item   S_X2 df.S_X2 RMSEA.S_X2 p.S_X2
  1    i1  2.821      3    0.000  0.420
  2    i2  1.884      5    0.000  0.865
  3    i3  6.021      5    0.014  0.304
  4    i4 10.093      5    0.032  0.073
  5    i5  6.305      5    0.016  0.278
  6    i6 11.977      5    0.037  0.035
  7    i7  6.046      5    0.014  0.302
  8    i8  1.760      5    0.000  0.881
  9    i9  6.874      5    0.019  0.230
 10   i10  6.420      5    0.017  0.267

> coef(M2PL2D,simplify=TRUE)
 $items
   a1     a2     d g u
 i1 2.293 0.753 1.668 0 1
 i2 1.096 0.301 0.304 0 1
 i3 1.286 0.396 0.649 0 1
 i4 0.717 0.690 0.548 0 1
 i5 0.706 0.350 0.057 0 1
 i6 0.808 1.205 -1.394 0 1
 i7 0.593 1.215 -0.857 0 1
 i8 0.472 1.049 -1.044 0 1
 i9 0.459 1.336 -1.723 0 1
 i10 0.789 0.756 -1.739 0 1

```

(continued)

TABLE 10.4. (continued)

```

$means
F1 F2
 0  0

$cov
      F1 F2
F1 1.000 NA
F2 0.158  1

> summary(M2PL2D)
      F1     F2     h2
i1  0.776 0.255 0.668
i2  0.536 0.147 0.309
i3  0.593 0.182 0.385
i4  0.364 0.350 0.255
i5  0.376 0.187 0.177
i6  0.361 0.539 0.421
i7  0.273 0.559 0.387
i8  0.230 0.510 0.313
i9  0.207 0.604 0.408
i10 0.390 0.374 0.292

SS loadings:  1.968 1.645
Proportion Var:  0.197 0.165

Factor correlations:

      F1     F2
F1 1.000 0.158
F2 0.158 1.000

> # Three dimensional solution =====
> ModelSpec3D=mirt.model('F1 = 1-10
+ F2 = 1-10
+ F3 = 1-10
+ COV = F1*F2*F3')

> print((M2PL3D = mirt(intprnsndat,ModelSpec3D,'2PL', method = 'MHRM',SE=T)))

Stage 3 = 126, LL = -8882.1, AR(1.00) = [0.32], gam = 0.0048, Max-Change = 0.0010

Calculating information matrix...

Calculating log-likelihood...

Call:
mirt(data = intprnsndat, model = ModelSpec3D, itemtype = "2PL",
SE = T, method = "MHRM")

Full-information item factor analysis with 3 factor(s).
Converged within 0.001 tolerance after 126 MHRM iterations.
mirt version: 1.30
M-step optimizer: NR1
Latent density type: Gaussian

Information matrix estimated with method: MHRM
Condition number of information matrix = 158.2028
Second-order test: model is not a maximum, or the information matrix is too inaccurate

Log-likelihood = -5686.907, SE = 0.021
Estimated parameters: 43
AIC = 11459.81; AICc = 11463.77
BIC = 11670.85; SABIC = 11534.28
G2 (980) = 694.39, p = 1
RMSEA = 0, CFI = NaN, TLI = NaN

```

(continued)

TABLE 10.4. (*continued*)

```

> M2 (M2PL3D,CI=0.95)
      M2 df          p RMSEA RMSEA_2.5  RMSEA_97.5      SRMSR       TLI CFI
stats 4.781773 12 0.9648734     0          0 0.009136072 0.01267897 1.01501    1

> # 2D vs 3D model-level fit comparison -----
> anova(M2PL2D,M2PL3D)
  Model 1: mirt(data = intprnsndat, model = ModelSpec2D, itemtype = "2PL",
                 SE = T, method = "MHRM")
  Model 2: mirt(data = intprnsndat, model = ModelSpec3D, itemtype = "2PL",
                 SE = T, method = "MHRM")

      AIC      AICc     SABIC      HQ      BIC      logLik      X2   df      p
1 11444.89 11446.94 11498.57 11502.71 11597.03 -5691.444    NaN  NaN  NaN
2 11459.82 11463.77 11534.28 11540.02 11670.85 -5686.907 9.074 12 0.697

> coef(M2PL3D,simplify=TRUE)
$items
      a1      a2      a3      d      g      u
i1  1.212  2.470  0.463  2.008  0 1
i2  0.253  0.855  0.488  0.307  0 1
i3  0.184  1.052  0.793  0.673  0 1
i4  0.361  0.378  0.886  0.575  0 1
i5  0.454  0.505  0.173  0.060  0 1
i6  1.243  0.258  0.568 -1.411  0 1
i7  1.667 -0.054  0.373 -0.954  0 1
i8  0.723  0.026  0.839 -1.052  0 1
i9  1.104 -0.015  0.645 -1.684  0 1
i10 0.491  0.416  0.775 -1.750  0 1

$means
F1 F2 F3
0 0 0

$cov
      F1      F2      F3
F1 1.000  NA  NA
F2 0.324  1.00  NA
F3 0.291  0.15  1

> summary(M2PL3D)
      F1      F2      F3      h2
i1  0.3708  0.75583 0.142  0.729
i2  0.1277  0.43135 0.246  0.263
i3  0.0853  0.48715 0.367  0.379
i4  0.1815  0.18988 0.446  0.268
i5  0.2469  0.27418 0.094  0.145
i6  0.5657  0.11754 0.258  0.401
i7  0.6911 -0.02233 0.155  0.502
i8  0.3559  0.01264 0.413  0.298
i9  0.5188 -0.00715 0.303  0.361
i10 0.2480  0.21030 0.392  0.259

SS loadings:  1.51 1.165 0.93
Proportion Var:  0.151 0.116 0.093

Factor correlations:

      F1      F2      F3
F1 1.000 0.324 0.291
F2 0.324 1.000 0.150
F3 0.291 0.150 1.000

```

(continued)

TABLE 10.4. (continued)

```
# > M2PL - 2D additional results
.....
> itemplot(M2PL2D,1,rot = list(xaxis = -70, yaxis = 40, zaxis = 10),
theta_lim=c(-4,4),degrees=20) # produces Figure 10.9

> plot(M2PL2D,type="infocontour",theta_lim=c(-8,8)) # produces Figure 10.10 (left)
> plot(M2PL2D,type="info",theta_lim=c(-6,6)) # produces Figure 10.10 (right)

> head((peopleM2PL2D$fscores(M2PL2D, "EAP", full.scores.SE = T)),6)
      F1        F2       SE_F1       SE_F2
[1,] 0.44800057 -0.70767684 0.6554987 0.7446564
[2,] 0.16186358 -0.45294176 0.6378614 0.7361369
[3,] 0.41705491 -0.50059872 0.6581129 0.7306350
[4,] -0.02708463  0.01746118 0.6339619 0.7096553
[5,] -0.04759402 -0.04711218 0.6310668 0.7145551
[6,] -0.68424167 -0.23738718 0.6105110 0.7513799

> tail(peopleM2PL2D,4)
      F1        F2       SE_F1       SE_F2
[997,] -0.9227519 -0.9080829 0.6385991 0.8117934
[998,] -0.1779810  0.5359047 0.6342885 0.6850367
[999,]  1.3970843  1.6569960 0.8005078 0.7271217
[1000,] -0.4251188 -0.5263048 0.6099206 0.7636514

> mean(peopleM2PL2D[,1]) # average person location estimate, dimension 1
[1] -0.003801367

> mean(peopleM2PL2D[,2]) # average person location estimate, dimension 2
[1] -0.001051065

> sd(peopleM2PL2D[,1]) # person location estimate variable, dimension 1
[1] 0.7455191

> sd(peopleM2PL2D[,2]) # person location estimate variable, dimension 2
[1] 0.686428

> write.csv(peopleM2PL2D, file = "peopleM2PL2D_EAP.csv")

> plot(M2PL2D,type="scorecontour",theta_lim=c(-4,4)) # Figure 10.11 (left)
> plot(M2PL2D,type="score",theta_lim=c(-4,4)) # Figure 10.11 (right)

> peopleM2PL2DFit=personfit(M2PL2D,method="EAP")

> # combine theta hats & fit info
> peopleM2PL2DnFit=cbind(peopleM2PL2D,peopleM2PL2DFit)
> head(peopleM2PL2DnFit,6)
      F1        F2       SE_F1       SE_F2       Zh
 1 0.44800057 -0.70767684 0.6554987 0.7446564  1.0798294
 2 0.16186358 -0.45294176 0.6378614 0.7361369  0.8868903
 3 0.41705491 -0.50059872 0.6581129 0.7306350 -0.1267980
 4 -0.02708463  0.01746118 0.6339619 0.7096553  0.4507465
 5 -0.04759402 -0.04711218 0.6310668 0.7145551  0.2820920
 6 -0.68424167 -0.23738718 0.6105110 0.7513799 -0.4663463

> tail(peopleM2PL2DnFit,4)
      F1        F2       SE_F1       SE_F2       Zh
 997 -0.9227519 -0.9080829 0.6385991 0.8117934 0.6794231
 998 -0.1779810  0.5359047 0.6342885 0.6850367 0.2241238
 999  1.3970843  1.6569960 0.8005078 0.7271217 1.1644664
1000 -0.4251188 -0.5263048 0.6099206 0.7636514 1.0164625
```

TABLE 10.5. flexMIRT Command File and Abridged Output MIRT Analysis

```

<Project>
Title = "interpersonal engagement";
Description = "10 Items, 2PL, case data";

<Options>
Mode = Calibration;
GOF = Complete;
NumDec = 3;
savePRM= Yes;
SCORE= EAP;
saveSCO= Yes;
FisherInf= 81, 4.0;
saveINF= Yes;
SE= Fisher;
MaxE = 1000;
Etol = 1e-4;
Mtol = 1e-5;

<Groups>
%OnlyGroup%
File = "intrprsnl.dat";                                // space delimited file
Varnames = i1-i10;
N = 1000;
Ncats(i1-i10)=2;
Model(i1-i10) = Graded(2);                            // 2PL model

<Constraints>

< Abridged 2PL model output>
:
Summary of the Data and Dimensions
  Missing data code      -9
  Number of Items        10
  Number of Cases        1000
# Latent Dimensions      1
:
Maximum number of cycles: 1000
Convergence criterion: 1.00e-004
Maximum number of M-step iterations: 100
Convergence criterion for iterative M-steps: 1.00e-005
Number of rectangular quadrature points: 49
Minimum, Maximum quadrature points: -6.00, 6.00
Standard error computation algorithm: Fisher (Expected)
:
Number of free parameters: 20
Number of cycles completed: 25
:
Processing times (in seconds)
:
Total: 0.12
:
Convergence and Numerical Stability
flexMIRT(R) engine status: Normal termination
First-order test: Convergence criteria satisfied
Condition number of information matrix: 13.2307
Second-order test: Solution is a possible local maximum
:

```

(continued)

item 6, all nonsignificant. (For more information on the above indices see Appendix G “CFI, GFI, M_2 , RMSEA, TLI, and SRMR.”)

Our item parameter estimates are shown in our coefficient table. For example, item 1 has $\hat{\alpha}_{11} = 2.293$, $\hat{\alpha}_{12} = 0.753$, and $\hat{\gamma}_1 = 1.668$; for item 2 we have $\hat{\alpha}_{21} = 1.096$, $\hat{\alpha}_{22} = 0.301$, and $\hat{\gamma}_2 = 0.304$; and so on. (By setting `printSE` to TRUE in our call to `coef` we obtain our items' standard errors.) Bearing in mind that our first dimension is `sirt.harm`'s second dimension and our second is `sirt.harm`'s first dimension, our estimates correlate 0.99 or higher with those of `sirt.harm`. The standardized covariance (i.e., Pearson correlation) between our two dimensions is estimated to be 0.158. By using `itemplot` we can obtain our item response surfaces; the surface can be rotated (`rot`) if desired. For example, Figure 10.9 shows item 1's IRS (`itemplot(M2PL2D, 1, . . .)`). Additionally, we can use the `plot` function to see the instrument's total information surface as a contour plot (`plot(. . . , type = "infocontour" . . .)`) and as a surface (`plot(. . . , type = "info", . . .)`). Figure 10.10 shows both variants. Our instrument's peak information (about 3.5) falls between approximately -0.6 and -1.8 on dimension 1 and 0.6 and 2.8 on dimension 2. The contour shows a steep drop-off as we move in directions at $\sim 45^\circ$ and at $\sim 225^\circ$ from the center of the peak. Perpendicular to this path, the total information decreases more slowly. The contour plot allows one to see the backside of the total information surface without having to rotate the surface.

As in Chapter 4, we obtain our person EAP estimates and their standard errors by using the `fscores` function (`fscores(M2PL2D, "EAP", full.scores.SE = T)`). Our first person is estimated to be located at 0.448 ($s_e(\hat{\theta}_1) = 0.655$) on dimension 1 and -0.708 ($s_e(\hat{\theta}_1) = 0.745$) on dimension 2, our second person's estimates are $\hat{\theta}_2 = 0.162$ ($s_e(\hat{\theta}_2) = 0.638$) and $\hat{\theta}_2 = -0.453$ ($s_e(\hat{\theta}_2) = 0.736$), etcetera. These estimates can be transformed to another scale, such as an expected trait score scale. For instance, Figure 10.11 shows the expected trait scores as contour and surface plots. Graphically, the transformation is easier to perform with the contour plot than with the surface plot. As can be seen from the contour plot, different $\hat{\theta}_j$'s can lead to the same expected trait score, T (i.e., a contour line); with 10 items our contour lines are bounded by 0 and $L = 10$. Alternatively, different θ_j 's will produce different T s conditional on, say, θ_1 .

Example: Calibration of Interpersonal Engagement Instrument, M2PL Model, flexMIRT

The flexMIRT command file for this calibration is presented in Table 10.5. To perform our unidimensional 2PL model calibration we use the relationship between the GR and 2PL models. Accordingly, we specify that each item has two categories (`Ncats(i1-i10) = 2`) and to apply the GR model to each of our ten items (`Model1(i1-i10) = Graded(2)`). For the multidimensional analyses we build upon the 2PL base program by adding the `FactorLoadings` command to its `Options` section and the `Rotations`, `Oblique`, and `Dimensions` statements to its `Groups` section. By setting `Dimensions` to 2 or 3 we specify two- or three-dimensional analyses, respectively.

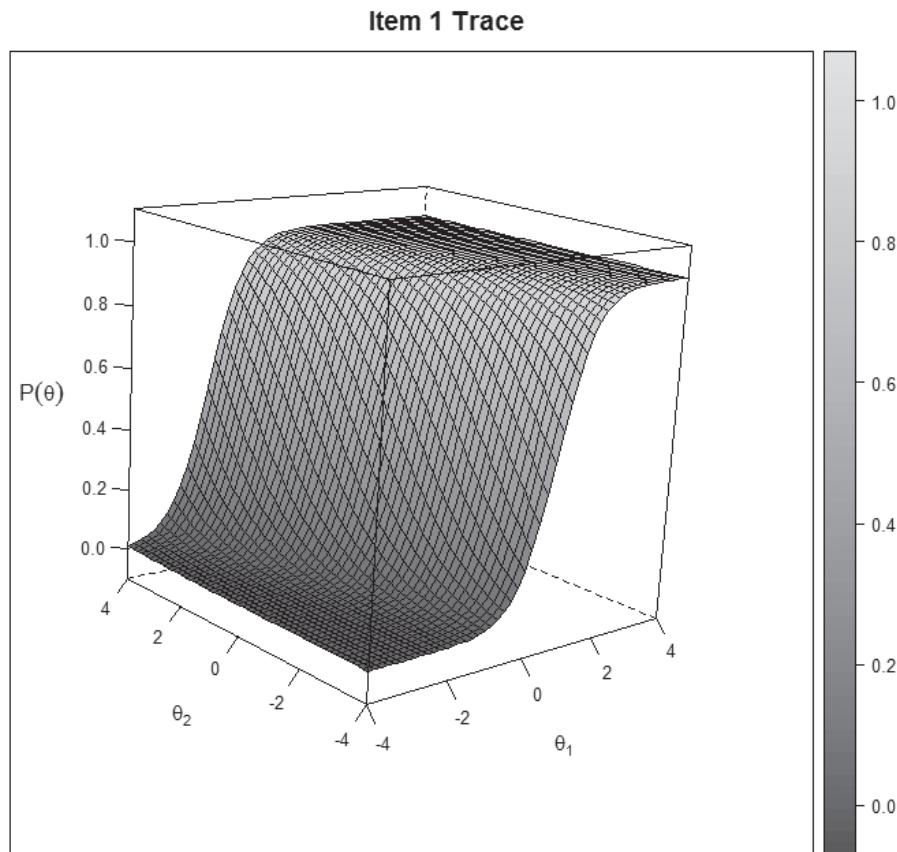


FIGURE 10.9. Item response surface for item 1 from the interpersonal engagement scale.

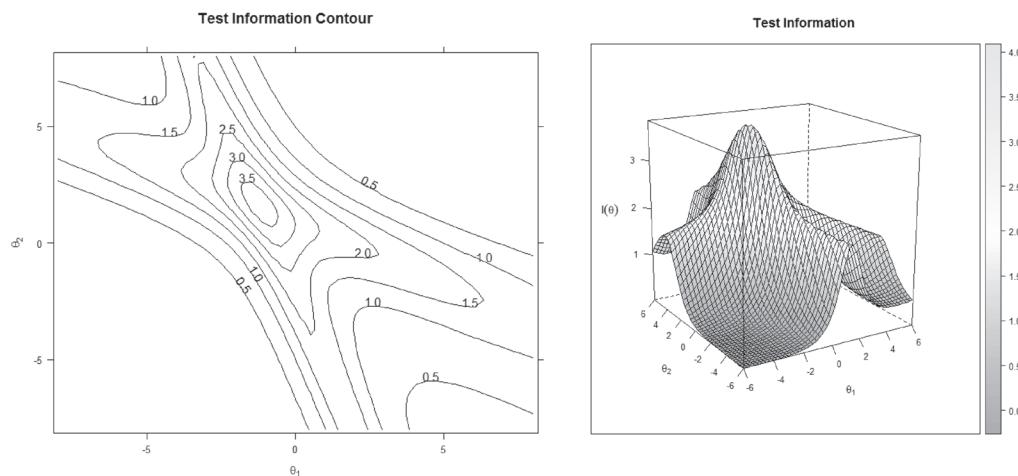


FIGURE 10.10. Total information contour (left) and surface plot (right).

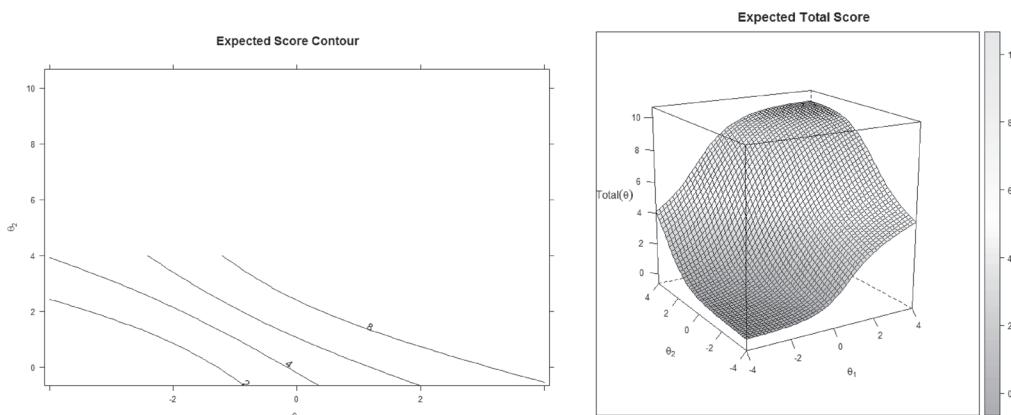


FIGURE 10.11. Total characteristic contour (left) and surface (right).

The bottom panel of Table 10.2 contains flexMIRT's model-level fit information. These results parallel those of mirt and sirt.noharm. Specifically, with respect to our information criteria, BIC indicates that the one-dimensional model is a better fit than either the two- or three-dimensional model, whereas our AIC and G^2 indicate that the two-dimensional model is a better fit than the one-dimensional model. Moreover, the additional complexity of the three-dimensional model over the two-dimensional model is not warranted.

Examining the 2PL model output, we see that item 1 has an estimated discrimination (a) of 2.023 and an estimated location (b) of -0.706 ($c = \hat{\gamma}_1 = 1.428$), for item 2 we have $\hat{a}_2 = 1.072$, $\hat{b}_2 = -0.266$ ($\hat{\gamma}_2 = 0.285$), and so on; the $S - X^2$ fit statistics are all nonsignificant. For our two-dimensional M2PL we see that $\alpha_{12} = 0.0$ ($a_{12} = 2$) for model identification, and there are some negative discrimination estimates on this dimension; the $S - X^2$ fit statistics are all nonsignificant. Our rotated discrimination estimates correlate between 0.986 and 0.999 with those of sirt.noharm and mirt, whereas our intercept estimates correlate 1.000 with sirt.noharm's and mirt's estimated intercepts.

Summary

Some data cannot be appropriately modeled by our unidimensional IRT models. In some cases, the (*between-item*) multidimensionality may be addressed by decomposing the instrument into multiple unidimensional components and utilizing a unidimensional model with each component. In other situations, it is not possible to decompose the instrument into unidimensional components because the responses are a manifestation of more than one latent continuous variable (i.e., *within-item multidimensionality*). Data of this type may be more appropriately modeled by using a multidimensional IRT model.

When more than one latent variable is necessary to respond to an item, there is

TABLE 10.5. flexMIRT Command File and Abridged Output MIRT Analysis

```

<Project>
Title = "interpersonal engagement";
Description = "10 Items, 2PL, case data";

<Options>
Mode = Calibration;
GOF = Complete;
NumDec = 3;
savePRM= Yes;
SCORE= EAP;
saveSCO= Yes;
FisherInf= 81, 4.0;
saveINF= Yes;
SE= Fisher;
MaxE = 1000;
Etol = 1e-4;
Mtol = 1e-5;

<Groups>
%OnlyGroup%
File = "intrprsnl.dat";                                // space delimited file
Varnames = i1-i10;
N = 1000;
Ncats(i1-i10)=2;
Model(i1-i10) = Graded(2);                            // 2PL model

<Constraints>

< Abridged 2PL model output>
:
Summary of the Data and Dimensions
  Missing data code      -9
  Number of Items        10
  Number of Cases        1000
# Latent Dimensions      1
:
Maximum number of cycles: 1000
Convergence criterion: 1.00e-004
Maximum number of M-step iterations: 100
Convergence criterion for iterative M-steps: 1.00e-005
Number of rectangular quadrature points: 49
Minimum, Maximum quadrature points: -6.00, 6.00
Standard error computation algorithm: Fisher (Expected)
:
Number of free parameters: 20
Number of cycles completed: 25
:
Processing times (in seconds)
:
Total: 0.12
:
Convergence and Numerical Stability
flexMIRT(R) engine status: Normal termination
First-order test: Convergence criteria satisfied
Condition number of information matrix: 13.2307
Second-order test: Solution is a possible local maximum
:

```

(continued)

TABLE 10.5. (continued)

2PL Items for Group 1: OnlyGroup

Item	Label	P#	a	s.e.	P#	c	s.e.	b	s.e.
1	i1	2	2.023	0.223	1	1.428	0.144	-0.706	0.064
2	i2	4	1.072	0.113	3	0.285	0.079	-0.266	0.076
3	i3	6	1.264	0.129	5	0.611	0.088	-0.483	0.074
4	i4	8	1.066	0.114	7	0.543	0.082	-0.509	0.084
5	i5	10	0.809	0.097	9	0.052	0.072	-0.064	0.090
6	i6	12	1.478	0.155	11	-1.363	0.115	0.922	0.085
7	i7	14	1.279	0.132	13	-0.815	0.092	0.638	0.079
8	i8	16	1.093	0.120	15	-1.005	0.091	0.919	0.103
9	i9	18	1.233	0.139	17	-1.595	0.114	1.294	0.121
10	i10	20	1.166	0.137	19	-1.738	0.117	1.491	0.143

:

Orlando-Thissen-Bjorner Summed-Score Based Item Diagnostic Tables and X2s:

Group 1: OnlyGroup

Item 1 S-X2(6) = 1.5, p = 0.9599
 Item 2 S-X2(8) = 4.1, p = 0.8479
 Item 3 S-X2(7) = 5.2, p = 0.6319
 Item 4 S-X2(8) = 11.7, p = 0.1643
 Item 5 S-X2(8) = 6.2, p = 0.6261
 Item 6 S-X2(8) = 13.5, p = 0.0945
 Item 7 S-X2(8) = 7.2, p = 0.5130
 Item 8 S-X2(8) = 5.6, p = 0.6874
 Item 9 S-X2(8) = 5.2, p = 0.7348
 Item 10 S-X2(8) = 6.1, p = 0.6373

Summed Score to Scale Score Conversion Table:

Summed

Score	EAP	SD	P	O
0.00	-1.629	0.636	0.0473446	0.0480000
1.00	-1.181	0.587	0.0890452	0.0830000

:

10.00 2.009 0.619 0.0140775 0.0130000

Summed score based latent distribution fit S-D2 = 6.8, p = 0.5629
 Marginal reliability of the scaled scores for summed scores = 0.70613

:

Statistics based on the loglikelihood of the fitted model:

-2loglikelihood: 11417.30
 Akaike Information Criterion (AIC): 11457.30
 Bayesian Information Criterion (BIC): 11555.45

Full-information fit statistics of the fitted model:

Degrees		G2 of freedom	Probability	F0hat	RMSEA
738.22	326	0.0001	0.7382	0.04	

Degrees		X2 of freedom	Probability	F0hat	RMSEA
1372.36	1003	0.0001	1.3724	0.02	

:

< Abridged M2PL model 2D command file & output>

:

<Options>

:

FactorLoadings=Yes;

<Groups>

:

Dimensions = 2; // two-dimensional solution
 Rotation=CFvarimax;
 Oblique=No;
 Model(i1-i10) = Graded(2); // M2PL model

(continued)

TABLE 10.5. (continued)

```

:
< output begins >
Maximum number of cycles: 1000
Convergence criterion: 1.00e-004
Maximum number of M-step iterations: 100
Convergence criterion for iterative M-steps: 1.00e-005
Number of rectangular quadrature points: 49
Minimum, Maximum quadrature points: -6.00, 6.00
Standard error computation algorithm: Fisher (Expected)
:
Number of free parameters: 29
Number of cycles completed: 154
:
Processing times (in seconds)
:
Total: 10.89
:
Convergence and Numerical Stability
flexMIRT(R) engine status: Normal termination
First-order test: Convergence criteria satisfied
Condition number of information matrix: 64.9266
Second-order test: Solution is a possible local maximum
:
2PL Items for Group 1: OnlyGroup
  Item      Label    P#   a 1   s.e.    P#   a 2   s.e.    P#     c   s.e.
    1        i1      2   2.460   0.412    0.000   ----  1   1.622   0.222
    2        i2      4   1.211   0.149    5   -0.060   0.203  3   0.299   0.083
    3        i3      7   1.416   0.174    8   -0.032   0.226  6   0.641   0.095
    4        i4     10   0.983   0.121   11   0.411   0.172  9   0.542   0.082
    5        i5     13   0.823   0.109   14   0.107   0.156 12   0.052   0.073
    6        i6     16   1.294   0.169   17   0.891   0.233 15   -1.406   0.124
    7        i7     19   1.113   0.159   20   1.033   0.248 18   -0.883   0.108
    8        i8     22   0.920   0.138   23   0.813   0.205 21   -1.049   0.100
    9        i9     25   1.027   0.165   26   1.035   0.258 24   -1.712   0.146
   10       i10    28   1.083   0.151   29   0.437   0.201 27   -1.741   0.117

Factor Loadings for Group 1: OnlyGroup
  Item      Label lambda 1   s.e. lambda 2   s.e.
    1        i1      0.822  0.076   0.000   ----
    2        i2      0.580  0.080   -0.029  0.165
    3        i3      0.639  0.079   -0.014  0.173
    4        i4      0.490  0.082   0.204  0.143
    5        i5      0.435  0.081   0.057  0.141
    6        i6      0.559  0.097   0.385  0.152
    7        i7      0.488  0.100   0.453  0.150
    8        i8      0.438  0.099   0.387  0.146
    9        i9      0.458  0.111   0.462  0.159
   10       i10     0.525  0.098   0.212  0.163

Orthogonal CF-Varimax Rotated Loadings for Group 1: OnlyGroup
  Item      Label lambda 1 lambda 2
    1        i1      -0.304  -0.764
    2        i2      -0.188  -0.549
    3        i3      -0.223  -0.599
    4        i4      -0.371  -0.379
    5        i5      -0.213  -0.383
    6        i6      -0.564  -0.377
    7        i7      -0.601  -0.286
    8        i8      -0.522  -0.264
    9        i9      -0.599  -0.255
   10       i10     -0.391  -0.409

```

(continued)

TABLE 10.5. (continued)

Orlando-Thissen-Bjorner Summed-Score Based Item Diagnostic Tables and X2s:
 Group 1: OnlyGroup
 Item 1 S-X2(7) = 4.1, p = 0.7682
 Item 2 S-X2(7) = 5.6, p = 0.5926
 Item 3 S-X2(6) = 7.1, p = 0.3106
 Item 4 S-X2(7) = 11.7, p = 0.1104
 Item 5 S-X2(7) = 6.1, p = 0.5344
 Item 6 S-X2(7) = 12.1, p = 0.0968
 Item 7 S-X2(7) = 6.9, p = 0.4365
 Item 8 S-X2(7) = 4.6, p = 0.7034
 Item 9 S-X2(7) = 6.8, p = 0.4487
 Item 10 S-X2(7) = 6.1, p = 0.5296

Summed Score to Scale Score Conversion Table:

Score	EAP 1	EAP 2	SD 1	SD 2	P	O	Error Covariance Matrix
0.00	-1.606	-0.334	0.637	0.923	0.0485558	0.0480000	0.405211
1.00	-1.165	-0.259	0.600	0.942	0.0880774	0.0830000	-0.085500 0.851885
							0.359482
							-0.126506 0.887544
10.00	1.759	0.987	0.705	0.862	0.0147909	0.0130000	0.497149
							-0.218376 0.742952

Summed score based latent distribution fit S-D2 = 5.7, p = 0.6774

Statistics based on the loglikelihood of the fitted model:
 -2loglikelihood: 11381.72
 Akaike Information Criterion (AIC): 11439.72
 Bayesian Information Criterion (BIC): 11582.05

Full-information fit statistics of the fitted model:

Degrees		G2 of freedom		Probability	F0hat	RMSEA
702.64	317	0.0001		0.7026	0.03	
Degrees		X2 of freedom		Probability	F0hat	RMSEA
945.62	994	0.8618		0.9456	0.00	

< Abridged M2PL model 3D command file & output>

<Options>
 FactorLoadings=Yes;

<Groups>
 Dimensions = 3; // three-dimensional solution
 Rotation=CFvarimax;
 Oblique=No;
 Model(i1-i10) = Graded(2); // M2PL model

< output begins >
 Maximum number of cycles: 3000
 Convergence criterion: 1.00e-004
 Maximum number of M-step iterations: 100
 Convergence criterion for iterative M-steps: 1.00e-005
 Number of rectangular quadrature points: 49
 Minimum, Maximum quadrature points: -6.00, 6.00
 Standard error computation algorithm: Fisher (Expected)

(continued)

TABLE 10.5. (continued)

:
Number of free parameters: 37
Number of cycles completed: 1274
:
Processing times (in seconds)
:
Total: 3541.59
:
Convergence and Numerical Stability
flexMIRT(R) engine status: Normal termination
First-order test: Convergence criteria satisfied
Condition number of information matrix: 2337.3308
Second-order test: Solution is a possible local maximum
:
Statistics based on the loglikelihood of the fitted model:
-2loglikelihood: 11370.39
Akaike Information Criterion (AIC): 11444.39
Bayesian Information Criterion (BIC): 11625.98
Full-information fit statistics of the fitted model:
Degrees
G2 of freedom Probability F0hat RMSEA
691.31 309 0.0001 0.6913 0.04
Degrees
X2 of freedom Probability F0hat RMSEA
925.30 986 0.9164 0.9253 0.00

the possibility that a person's location on one latent variable may compensate for their location on other latent variable(s). Data of this type are modeled using a compensatory MIRT model. In contrast, when a person's location on one latent variable does not compensate for their location on other latent variable(s), then one has a noncompensatory/partial compensatory situation. Although noncompensatory/partial compensatory models exist, these models have not seen as much attention as the compensatory models. In addition to the models discussed in this chapter, there are a number of other models for multidimensional data (e.g., Bock & Aitkin, 1981; Embretson, 1997; Fischer & Seliger, 1997; Sympson, 1978; Whitely, 1980; also see Mislevy, 1986b). This chapter has focused on the compensatory multidimensional extensions of the 2PL and 3PL models.

The M2PL model is applicable to compensatory multidimensional data. This model is typically presented in a linearized form with an item discrimination parameter for each dimension and an intercept parameter, γ_j . As is true with the unidimensional models, one cannot directly interpret γ_j as a location parameter. To provide an interpretation analogous to a unidimensional model's location parameter, we calculate the item's multidimensional item location, Δ_j . An item's multidimensional item location indicates the distance from the origin in the latent space to the item's point of maximum discrimination in a particular direction from the origin. With the M2PL this point of maximum discrimination occurs when $p_j = 0.5$.

Although the discrimination parameters can be interpreted as in the unidimensional models, sometimes it is useful to have a single value that represents an item's discrimination power in the multidimensional space. This value is reflected in the item's

multidimensional discrimination parameter, A_j . The larger the value of A_j , the greater item j 's discrimination capacity across the F -dimensions. Moreover, analogous to the unidimensional models, the greater the item's discrimination capability, the greater the item's multidimensional information in a particular direction.

MIRT models have two sources of indeterminacy that need to be addressed when estimating parameters. As is the case with the unidimensional models, our first source of indeterminacy is the indeterminacy of the metric. The second source of indeterminacy is rotational indeterminacy. How these indeterminacies are addressed depends on the calibration program.

Graphical representation of a MIRT model's functional form is typically done by using an item response surface plot and/or contour plot. With a contour plot the contours represent the points of equal probability. Vector plots are typically used to display item characteristics for item sets. Vector plots are used to simultaneously present not only an item's location and how well it discriminates (i.e., Δ_j and A_j), but also which dimension, if any, it measures best.^{13,14}

The preceding chapters have been devoted to presenting IRT models for different types of data. As part of this presentation we have made comparisons with CTT and pointed out the advantages of IRT models over the true score model. One of these IRT advantages is the model's capacity to predict how individuals located at a particular point on a continuum would be expected to respond. Consequently, we can administer only those items that provide the most information for estimating an individual's location. This is the premise of computerized adaptive testing and is discussed in Appendix D. A second advantage, given that one has model–data fit, is the invariance of the parameter estimates. This property facilitates the creation of item banks with desirable characteristics as well as multiple forms of an instrument. The creation of multiple forms is useful for maintaining the veracity of person location estimates and for tracking changes in performance over time. The creation of item pools and multiple forms requires that one be able to link the various instrument administrations onto a common metric. The next chapter addresses the creation of a common metric across multiple forms and/or samples of individuals.

Notes

1. An example of a noncompensatory or partially compensatory is Sympson's (1978) model. In his multidimensional 3PL model, the probability of a response of 1 given person i 's and item j 's locations in the F -dimensional latent space is

$$p_j = p_j(x_j = 1 | \underline{\theta}_i, \underline{\alpha}_j, \underline{\delta}_j, \gamma_j) = \gamma_j + (1 - \gamma_j) \left(\prod_{f=1}^F p_j^*(\theta_{if}) \right), \quad (10.27)$$

where $\underline{\theta}_i$ is person i 's location vector (i.e., $\underline{\theta}'_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{iF})$), $\underline{\alpha}_j$ (i.e., $\underline{\alpha}'_j = (\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jF})$) and $\underline{\delta}_j$ (i.e., $\underline{\delta}'_j = (\delta_{j1}, \delta_{j2}, \dots, \delta_{jF})$) are item j 's discrimination and location vectors, respectively, γ_j is its pseudo-guessing parameter, and $p_j^*(\theta_{if})$ is essentially the 2PL model dimension-wise

$$p_j^*(\theta_{ij}) = \frac{\exp[\alpha_{jf}(\theta_{if} - \delta_{if})]}{1 + \exp[\alpha_{jf}(\theta_{if} - \delta_{if})]}.$$

As a consequence, item j has both a discrimination and location parameter for each dimension f . Because Equation 10.27 involves a product, the smallest $p_j^*(\theta_{if})$ affects p_j 's magnitude. In the extreme case where for dimension f we have $p_j^*(\theta_{if}) = 0$, then person i 's location(s) on the other dimension(s) cannot compensate for their location on dimension f regardless of their location(s) on the other dimensions; that is, p_j must equal 0. Because of this characteristic this model can be considered to be noncompensatory. However, when $p_j^*(\theta_{if}) > 0$ for dimension f , then person i 's location(s) on the other dimension(s) can somewhat affect p_j 's value. From this perspective this model can be considered to be partially compensatory.

When $\gamma_j = 0$, we have a partially compensatory/noncompensatory multidimensional 2PL model. `mirt` will estimate the partially compensatory 2PL and 3PL models.

2. The bifactor (or bi-factor; Gibbons & Hedeker, 1992) model is mentioned in Chapter 6 Endnote 7 in connection with the testlet model (i.e., the testlet and bifactor models are equivalent [Li, Bolt, & Fu, 2006; Rijmen, 2010]). With the bifactor model, one has a general θ as well as additional θ s associated with item clusters. Consequently, in an F -dimensional space item responses are manifestations of the primary θ and one of the $F - 1$ item cluster θ s (i.e., items can only load on the primary θ and one of the $F - 1$ θ s). Figure 10.12 presents the measurement models for an example bifactor model and a two-dimension compensatory model. In the bifactor model we have a general θ (e.g., mathematical ability) that affects performance on our eight item instrument. In addition, the first four items are also affected by an additional latent variable (θ_1). Similarly, the last four items are affected by a different additional latent variable (θ_2). Such a scenario could arise if the first four items share some common content (e.g., a graph to which each item refers) and the last four items share content with one another (e.g., a passage to which each item refers). In contrast, with the two-dimensional compensatory model, each item is affected by both latent variables. Obviously, higher compensatory models are possible so that each item is affected by multiple latent variables. Bonifay and Cai (2017) discuss model fit in the context of multidimensionality. `flexMIRT` and `mirt` will estimate the bifactor model.
3. A number of model misspecification studies have examined the use of a unidimensional IRT model when the model should contain two dimensions. In general, when the true model is a compensatory two-dimensional IRT model, then the unidimensional $\hat{\delta}_j$ is found to be an estimate of the average of the δ s across the dimensions, the unidimensional $\hat{\alpha}_j$ is an estimate of $(\alpha_{j1} + \alpha_{j2})$, and the estimated person location is an average of the true θ s across the dimensions. In contrast, when the true model is a noncompensatory model, then the $\hat{\delta}_j$ is an overestimate of or correlated more highly with one dimension δ_{if} than with the other, δ_{jf} , $\hat{\alpha}_j$ is an estimate of the average of the true α s, and $\hat{\theta}$ is an estimate of the average true θ s (Ackerman, 1989; Way,

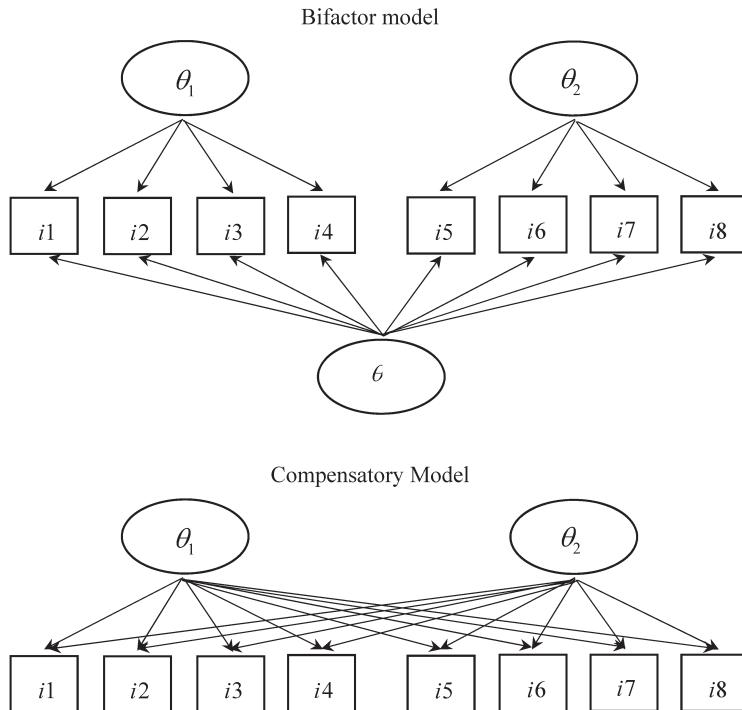


FIGURE 10.12. Graphical depiction of compensatory and bifactor models for two-dimensional situation.

Ansley, & Forsyth, 1988; Reckase, 1979). Wang (1986, 1987) analytically determined that the “unidimensional estimates of item parameters are obtained with reference to a weighted composite of underlying latent traits” where the weights are “primarily a function of the discrimination vectors for the items, the correlations among the latent traits and, to a lesser extent, the difficulty parameters of the items” (Wang, 1987, p. 3). In short, the parameter estimates reflect a reference composite or trait (Wang, 1987).

4. An item’s multidimensional item location is typically symbolized as D_j , MID_j , or B_j (e.g., Reckase, 1985, 2009; Reckase & McKinley, 1991). For consistency with our convention of using Greek letters for parameters, the uppercase delta, Δ , is used to represent an item j ’s multidimensional item location (Δ_j).
5. An item’s multidimensional discrimination parameter is typically symbolized as $MDISC_j$ (e.g., Reckase & McKinley, 1991; Reckase, 1986), but in keeping with our convention of using Greek letters for parameters, the uppercase alpha, A , is used to symbolize item j ’s multidimensional discrimination parameter (A_j).
6. From this item vector perspective, we see that Equation 10.12 is the application of the Pythagorean theorem to determining a vector’s length. That is, the vector is the hypotenuse of a right triangle. As such, an item’s multidimensional discrimination

parameter is analogous to the square root of the item's communality from a principal axis analysis.

7. As is the case with the 2PL and 3PL models, there are normal ogival forms of the M2PL and M3PL models (e.g., see Bock & Aitkin, 1981; Mislevy, 1986b; Samejima, 1974).
8. This approach assumes that $\underline{\theta}$ is multivariate normal with mean $\underline{0}$ and covariance matrix \underline{I} . This is an additional assumption to those presented above. Because the MMLE implementation uses the frequencies of response patterns and the pairwise proportions of 1s and 0s, the approach is sometimes called full-information factor analysis (FIFA).
9. In some cases, one may need to attend to whether a program's estimates are on the normal metric.
10. An item's multidimensional information is typically symbolized as $MINF$ (e.g., Reckase & McKinley, 1991), but in keeping with our convention of using "I" to symbolize the concept of information, we symbolize item j 's multidimensional information as $I_{j\omega}(\underline{\theta})$
11. The axes' rotation is accomplished by using a transformation matrix, \underline{T} . Specifically, the discrimination parameters are transformed (Hirsch, 1989) by

$$\underline{\alpha}_j^* = \underline{\alpha}_j \underline{T} \quad (10.28)$$

and the person locations by

$$\underline{\theta}^* = (\underline{T}^{-1}) \underline{\theta}, \quad (10.29)$$

where \underline{T}^{-1} is the inverse of \underline{T} . For our example the transformation matrix is

$$\underline{T} = \begin{bmatrix} \cos \angle_{11^*} & \cos \angle_{12^*} \\ \cos \angle_{21^*} & \cos \angle_{22^*} \end{bmatrix}, \quad (10.30)$$

where $\cos \angle_{11^*}$ is the cosine of the angle of rotation between the unrotated θ_1 and its rotated position θ_1^* (e.g., 25°) and $\cos \angle_{22^*}$ is the cosine of the angle of rotation between the unrotated θ_2 and its rotated position θ_2^* .

Equations 10.28 and 10.29 show why the rotational transformation leaves the logit unchanged. Following Hirsch (1989), the substitution of Equations 10.28 and 10.29 into the exponent of the model in Equation 10.3 yields

$$\underline{\alpha}_j^* \underline{\theta}^* + \gamma_j = \underline{\alpha}_j' \underline{T} (\underline{T}^{-1}) \underline{\theta} + \gamma_j = \underline{\alpha}_j' \underline{\theta} + \gamma_j$$

because $\underline{T}(\underline{T}^{-1}) = \underline{I}$.

12. For comparison, we demonstrate the estimation of a noncompensatory/partial compensatory two-dimensional 2PL model to these data. Although our model specification is the same as the compensatory model (see Table 10.4), we change itemtype to be PC2PL:

```

> print((PcompM2PL2D = mirt(intprnsndat, ModelSpec2D,
+ itemtype="PC2PL", method = 'SEM')))
Stage 2 = 100, LL = -7054.0, AR(2.20) = [0.15], Max-Change = 0.2500

Calculating log-likelihood...

Call:
mirt(data = intprnsndat, model = ModelSpec2D, itemtype = "PC2PL",
      method = "SEM")

Full-information item factor analysis with 2 factor(s).
Converged within NA tolerance after 100 SEM iterations.
mirt version: 1.30
M-step optimizer: NR1
Latent density type: Gaussian

Log-likelihood = -5753.647
Estimated parameters: 61
AIC = 11629.29; AICC = 11637.36
BIC = 11928.67; SABIC = 11734.93
G2 (962) = 827.73, p = 0.9993
RMSEA = 0, CFI = NaN, TLI = NaN
Calculating log-likelihood..

> coef(PcompM2PL2D,simplify=TRUE)
$items
      a1     a2     a3     d1     d2     d3 g u
i1  1.480  2.579  4.037  5.641  1.897  3.187 0 1
i2  3.750  1.011  0.747  6.724  0.597  2.297 0 1
i3 -0.788  1.662  1.825  3.682  0.842 15.714 0 1
i4  1.105  0.704  2.325  1.037  1.928 15.921 0 1
i5  0.754  9.169 -3.910  0.381 12.542  7.912 0 1
i6 18.400 -5.856  1.702 17.132  7.781 -1.120 0 1
i7  1.478  0.494 -2.574 -0.547  2.984  7.814 0 1
i8  1.447 -1.919  0.043 -0.842  4.378  3.514 0 1
i9  2.060 -1.177  0.680 -1.222  1.906  3.439 0 1
i10 19.488 -1.280  5.811 14.722  3.106 -3.584 0 1

$means
F1 F2
0 0

$cov
      F1 F2
F1 1.0 NA
F2 0.8 1

> summary(PcompM2PL2D)
      F1        F2  (F1*F2)       h2
i1  0.2794  0.4871  0.7624  0.897
i2  0.8709  0.2348  0.1736  0.844
i3 -0.2541  0.5362  0.5886  0.699
i4  0.3490  0.2225  0.7346  0.711
i5  0.0744  0.9042 -0.3856  0.972

```

```
i6  0.9456 -0.3009  0.0875  0.992
i7  0.4276  0.1429 -0.7445  0.758
i8  0.4912 -0.6517  0.0146  0.666
i9  0.6872 -0.3925  0.2269  0.678
i10 0.9531 -0.0626  0.2842  0.993
```

SS loadings: 3.727 2.141 2.341
Proportion Var: 0.373 0.214 0.234

Factor correlations:

	F1	F2
F1	1.0	0.8
F2	0.8	1.0

Comparing this model's AIC and BIC values to the compensatory M2PL model shows the latter model fits better than the former model.

13. As Rost (1990) indicates, “multidimensionality is often seen as the only way out of the inaccuracies of unidimensional models” (p. 281). Therefore, as a counterpoint to the preceding models, we present mixture models in Appendix F, “Mixture Models.” These models may be useful in some multidimensional situations.
14. Reckase et al. (1988) show how a multidimensional calibration can be used to find clusters of items that measure the same weighted composite of traits that *meet* the assumption of unidimensionality. Therefore, it is possible to apply a unidimensional model to these item clusters. Their approach exploits the fact that “any item that can be described by the M2PL model is unidimensional in that it is equivalent to an item described by a unidimensional model, with the <latent variable> scale equal to a weighted composite of the elements of the θ -vector” (p. 195).

11

Linking and Equating

In the preceding chapters we have situations where we want to directly compare the item parameter estimates from different samples or models. However, because these estimates are on different metrics, they could not be directly compared with one another. The process of aligning different metrics is known as *linking*. When we align metrics to compare person location estimates, the process is called *equating*. In this chapter we begin by discussing the general process, including data collection approaches, and then we present different methods for transforming different metrics to a common metric.

Equating Defined

Equating refers to a family of procedures that transform person location estimates on different metrics to a common metric.¹ Therefore, the purpose of equating is to facilitate comparing individuals. For instance, we may be interested in assessing change over a year. As part of our study, we assess our participants quarterly, but to minimize carry-over effects we administer multiple forms of our instrument. In practice it is impossible, in terms of statistical characteristics, to create precisely identical forms. In addition, differences in administration conditions may affect the individuals' performances on the forms and thereby affect the forms' statistical characteristics. These slight differences in forms affect the person location estimates (e.g., X or $\hat{\theta}$). By equating the person location estimates across forms, we can (reasonably) eliminate the forms' differences to compare individuals. If an equating is successful, then it should not in principle make any difference which forms are administered to individuals (Lord, 1980). This is the overarching goal of equating.

Livingston (2004) provides a simple general definition of *equating*: A score on a new form and a score on another form are equivalent in a group of individuals that have taken the form if they represent the same *relative position* in the group. How "relative position" is operationalized differs across equating methods. For example, assume that

we have two alternate forms of an instrument. Some equating approaches (e.g., *mean equating*) would consider the scores on the two forms to be equated when the adjustment makes the mean of the scores on one form equal to the mean score on the other form. Therefore, in mean equating, the relative position of equated scores is defined in terms of the number of points the scores are from their respective means (Livingston, 2004). In other cases, such as *linear equating*, equating concerns itself with making the mean and standard deviation of, for example, the observed scores, on one form equal to those on the other form. Thus, in linear equating, the “relative position” of equated scores is defined in terms of the number of standard deviation units the scores are from their respective means. Stated another way, in mean equating the observed scores on one form are assumed to differ by a constant amount from those on the other form, whereas in linear equating the observed scores on one form may differ from those on the other form by different amounts, depending on their location on the observed score scale. A third strategy, *equipercentile equating*, operationalizes “relative position” in terms of the observed scores’ percentile ranks on the two forms. Specifically, scores are considered equated across the two forms if they have the same percentile rank on both forms. When equipercentile equating is successful, both instruments’ distributions have approximately the same mean, standard deviation, and distributional shape; equipercentile equating does not assume a uniform difference in difficulty between two test forms across the score scale (Kolen & Brennan, 2004). A general equating approach that subsumes equipercentile equating and so on, is kernel equating (KE; Holland & Thayer, 1989; von Davier, Holland, & Thayer, 2004); KE is discussed in Appendix G, “An Introduction to Kernel Equating.”

This chapter focuses on IRT equating and linking. The reader is referred to Angoff (1984), Holland and Rubin (1982), and Kolen and Brennan (1995, 2004) for detailed information on the traditional equating methods and their variants. Although these procedures may be performed using a statistical package (e.g., equipercentile equating with cubic spline smoothing may be performed using SAS), specialized programs for performing traditional equating are available. For instance, to perform equipercentile equating, one might use RAGE and RGEQUATE (Zeng, Kolen, Hanson, Cui, & Chien, 2004), the R packages *equate* (Albano, 2016, 2018) and *SNSEquate* (Gonzalez, 2014, 2020).

Implicit in equating is that the multiple forms are all measuring the same construct and that these forms have been created to the same content and statistical specifications (Angoff, 1984; Kolen & Brennan, 1995). As a contrarian example, one should not equate the scores on an algebra test to those on an art history test or on an androgyny scale. In addition, the transformation should be independent of the groups of individuals used to develop the transformation (i.e., the transformation is unique; Angoff, 1984).

The equating process itself may be seen as consisting of a data collection phase followed by a transformation phase. The transformation phase would involve the application of an equating technique (e.g., equipercentile equating, total characteristic function or characteristic function equating). We discuss these two phases in order.

Equating: Data Collection Phase

There are multiple data collection approaches or *equating designs*. Some of these strategies use a single sample (group) of individuals, whereas others use two samples. One of the single sample methods is referred to as the *single group with counterbalancing* (e.g., Kolen & Brennan, 1995) or the *counterbalanced random-groups design* (Petersen, Kolen, & Hoover, 1989). In this approach, we administer the two forms of an instrument to a single group of individuals. The administration uses counterbalancing of the form administration to control for order effects (e.g., fatigue, practice effects). For example, the single group would be decomposed into two subgroups. Subgroup 1 would receive form 1 followed by form 2 (form set 1), whereas subgroup 2 would receive form 2 followed by form 1 (form set 2). In practice, the counterbalancing is augmented by interleaving the form sets for administration. This interleaving is called *spiraling* and results in the first individual receiving form set 1, the second individual receiving form set 2, the third individual receiving form set 1, and so on (Angoff, 1984; Petersen et al., 1989). Because the same individuals are administered both forms, there is common information across the forms that can be used for performing the equating. In some situations it may not be practical to use this data collection strategy because it requires a doubling of administration time.

A second single-group design creates *randomly equivalent groups* (also known as *random groups* or *random subgroups*) from the single sample. In this approach half of the group is randomly assigned one form and the remaining half is assigned the other form. Spiraling is used to administer the forms (i.e., the first person receives form 1, the second person receives form 2, the third person receives form 1, and so on). This approach has the advantage of requiring half the administration time of the “single group with counterbalancing” design because each individual takes only one form. However, one has less common information for performing the equating than with the single group with counterbalancing. For both the “randomly equivalent groups” and the “single group with counterbalancing” designs it is necessary that both forms be available at the same time. As defined, the common information across forms is assumed to exist owing to the application of random assignment. A variant of this data collection design involves constructing the forms to have some items in common. This design, known as the *common-item random groups design* (Kolen & Brennan, 1995, 2004), uses the common items across the two forms to provide additional information for performing the equating.

A third data collection strategy, the *common-item nonequivalent groups design*, involves using two samples of individuals with each sample being administered one of the forms; this is also called the *nonequivalent groups with anchor test* (NEAT) design. As the name implies, the forms are created to contain some items in common. These common items are also referred to as *anchor items*, as a *common test*, or as an *anchor test* (Angoff, 1984; Kolen & Brennan, 1995, 2004; Lord, 1980; Petersen et al., 1989). The information used for equating the forms comes from the common items.

There are two variants of the common-item nonequivalent groups design. The first variant, *internal common items*, includes the performance on the common items as part

of the observed score. In the second variant, *external common items*, the individuals' performance on the common items is not considered part of their observed scores. With internal common items, the items are typically distributed throughout the instrument in the same general locations on both forms, whereas with external common items, they are typically presented after the noncommon items are administered (Lord, 1980). When administering the external common items in this way, fatigue, speededness, motivation, learning, practice, and so on, may have more of an impact on the equating than when using the internal common item approach.

The term *anchor test* implies how one should consider these common items. The common items should form a "miniversion" of the test or instrument (Klein & Jarjoura, 1985). That is, these items should be measuring the same construct, the same content specifications, and the same contextual effects as the noncommon items on the instrument (Angoff, 1984; Klein & Jarjoura, 1985; Kim & Cohen, 1998; Kolen & Brennan, 1995, 2004). An additional criterion for anchor tests to satisfy is that they have the same range of item locations as the total test. However, there may be some latitude with this criterion. For example, Sinharay and Holland (2007) found that anchor tests that had a range smaller than the total test, but that had been appropriately centered, performed as well as anchor tests that had the same item location range as the total test; also see Sinharay, Haberman, Holland, and Lewis (2012) and Ricker and von Davier (2007). Because their study used external common items and equipercentile equating, their results may not generalize to other equating procedures and/or the use of internal common items. An additional consideration for anchor tests in the context of IRT is that the model's assumptions be tenable for the common items and that these items not exhibit differential item functioning (Chapter 12). Typically, the anchor test size should be at least 20 items or no less than 20% of the instrument length (whichever is larger) (Angoff, 1984). In contrast to the single-group designs, this two-group strategy does not require that both forms be available for administration at the same time.²

Equating: Transformation Phase

The transformation phase of the equating process involves the application of one of the transformation procedures identified above (e.g., calculation of percentile ranks in equipercentile equating) to the collected response data. In the context of IRT this phase consists of potentially two steps. Assuming that one does not have item parameter estimates, then the first step involves obtaining model–data fit as well as the item parameter estimates for the administrations and the linking of the metrics. The second step is the actual application of the equating method to the person location estimates. Each of these is treated in turn below.

Recall that because of the latent variable's metric indeterminacy, the continuum's metric needs to be defined in the estimation process. This metric definition is addressed by the calibration program's centering strategy, as well as the calibration model, and is unique to a calibration sample. Therefore, if we administer the same instrument to two different groups, then each group defines its own metric for the latent continuum.³ As

mentioned above, when one has model–data fit, these two metrics are linearly related. This linear relationship is used to align metrics to form a common metric, and the alignment process is referred to as linking (cf. Koch & Reckase, 1979; Ree & Jensen, 1983).

Linking may be performed in a number of ways. For example, when one has administered different forms to different groups, one can perform *concurrent calibration* (also known as *simultaneous calibration*). In this approach, both samples' response data are concatenated and calibrated in one analysis. The result is that all the items are on the same metric—the one defined by all the individuals combined. If these item parameter estimates are then used to estimate the person locations, then the person locations are all on the same metric regardless of which form the person took (i.e., the individuals are equated). Figure 11.1 shows a visual representation of the structure of the input data file. For presentation purposes, all the common items in form 1 are shown as falling at the end of the form, and all the common items in form 2 are depicted as falling at the beginning of the form. The vertical aspect of the figure represents the two calibration samples—one for form 1 and another for form 2. The horizontal aspect reflects the items on the two forms. If we visually inspected the data file, we see a series of responses reflecting the responses to form 1 followed by blank spaces (or not-presented codes). Conversely, if one scrolled down through the file, one would find blank spaces (or not-presented codes) below form 1's responses, but upon scrolling across one would encounter form 2's responses. The simultaneous calibration strategy can be used with any of the equating designs mentioned above. In some contexts, this strategy may not be practical because of the number of samples and/or their size or because one wishes to maintain a particular baseline metric. Moreover, if the samples' latent distributions differ dramatically, then specifying the population parameters in certain estimation situations is difficult (Kim & Cohen, 1998).

A second approach makes use of previously obtained item parameter estimates for one of the forms. When the data for the second form are calibrated, the item parameter estimates from the first form are provided as input to the calibration and are held fixed (i.e., not re-estimated). In effect, this *fixed-item parameter* approach augments the second form's calibration by the first form's item calibration results. As a result, the second form's metric is aligned with the first form's metric. If these item parameter estimates are then used to estimate the person locations, then the person locations are on the same metric and the individuals are equated.

A third strategy involves using *metric transformation coefficients*. As mentioned in previous chapters, the continuum is determined up to a linear transformation. In general, the linear transformation from one metric to another for person and item locations (or their estimates) is

$$\xi^* = \zeta(\xi) + \kappa, \quad (11.1)$$

where ζ and κ are the unit and location coefficients, respectively. Typically, ζ and κ are referred to as *equating coefficients*. In this book they are referred to as metric transformation coefficients because they may or may not be used for equating. The symbol ξ represents the parameter (or its estimate) on the untransformed or *initial metric*, and

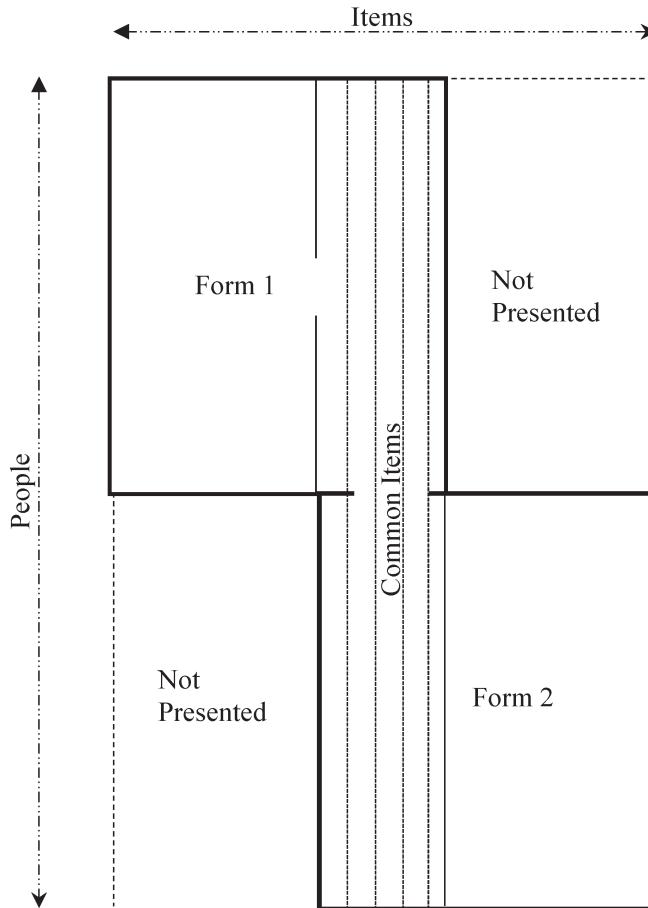


FIGURE 11.1. Graphical depiction of input data file for simultaneous calibration.

ξ^* represents the same parameter (or its estimate) transformed to the *target metric*. The target metric (a.k.a., common metric) is the metric onto which all other metrics are transformed.

For linking metrics ξ represents δ_j (or $\hat{\delta}_j$) on the initial metric and ξ^* is δ_j^* (or $\hat{\delta}_j^*$) on the target metric. For instance, by substitution into Equation 11.1 we have

$$\delta_j^* = \zeta(\delta_j) + \kappa. \quad (11.2)$$

To transform the initial metric's item discrimination parameter, α_j , to the target item discrimination parameter metric, α_j^* , we use

$$\alpha_j^* = \frac{\alpha_j}{\zeta} \quad (11.3)$$

or in the slope–intercept parameterization

$$\gamma_j^* = \gamma_j - \frac{\alpha_j(\kappa)}{\zeta}. \quad (11.4)$$

Equations 11.2–11.4 may also be applied to location, discrimination, and intercept parameters' estimates, respectively. The IRS's lower asymptote parameter, χ_j (or its estimate), are on a common [0, 1] metric and does not need to be transformed. To equate the person locations (or their estimates) Equation 11.1 is used, with ξ representing the initial metric person locations (or their estimates) and ξ^* reflecting the target metric person locations (or their estimates).

Multiple approaches for determining the values of ζ and κ have been developed using the common items. These approaches can be categorized as based on moments or on characteristic functions.

The simplest moments-based approach simply uses the means of the item parameters (or estimates) to obtain the metric transformation coefficients. Specifically, in the *mean-mean* approach (Loyd & Hoover, 1980) the equating coefficient ζ is given by

$$\zeta = \frac{\bar{\alpha}^*}{\bar{\alpha}}, \quad (11.5)$$

where $\bar{\alpha}^*$ and $\bar{\alpha}$ are the means of the common item discriminations on the target and initial metrics, respectively. Once ζ is determined, the other equating coefficient, κ , is obtained by

$$\kappa = \bar{\delta}^* - \zeta \bar{\delta}, \quad (11.6)$$

where $\bar{\delta}^*$ and $\bar{\delta}$ are the means of the common item locations (or their estimates) on the target and initial metrics, respectively.

Because the mean–mean approach ignores variability in the parameters (or estimates), it is most applicable when item discrimination does not vary. A second method, the *mean-sigma* method (Marco, 1977; also see Ree & Jensen, 1983), obtains the metric transformation coefficient ζ by using the standard deviations of the common items' locations. That is, ζ is the ratio of the target to initial metric standard deviations (s) of the locations

$$\zeta = \frac{s_{\delta}^*}{s_{\delta}}, \quad (11.7)$$

where s_{δ}^* is the standard deviation of the item locations (or their estimates) on the target metric and s_{δ} is the standard deviation of the item locations (or their estimates) on the initial metric. The κ metric transformation coefficient is obtained by Equation 11.6.

Once the metric transformation coefficients are obtained, then the linking of the separate metrics is performed by applying Equations 11.2 and 11.3 (or Equation 11.4) itemwise (or category/intersection-wise in the case of polytomous models) to the item parameter estimates. To equate the person locations across the metrics, we apply $\theta_i^* = \zeta(\theta_1) + \kappa$ to each individual's person location or its estimate. Either a program written in SAS, SPSS, SYSTAT (2017), etcetera, or specialized software (e.g., the ST program by Hanson and Zeng, 2004) or the R package SNSequate (Gonzalez, 2014, 2020) can be

used to implement this approach. This is the approach used in Chapter 4 and is applied to temperatures in Appendix G, “Linking: A Temperature Analogy Example.”

In contrast to only using the item parameter estimates’ moments, the *total characteristic function equating* uses all the item parameter estimates to determine the values of ζ and κ . The objective in this method (also known as *test characteristic curve equating*) is to align as closely as possible the initial metric’s total characteristic function (TCF) with that of the target metric. The metric transformation coefficients are the values of ζ and κ that satisfy this objective.

One variant of the characteristic function equating method was proposed by Haebara (1980) and another by Stocking and Lord (1983). Example programs that implement these approaches are EQUATE (Baker, 1993b; Stocking and Lord approach only), ST (Hanson & Zeng, 2004), POLYST (Kim & Kolen, 2003). The R packages *equateIRT* (Battauz, 2013, 2015, 2018), *plink* (Weeks, 2010, 2017), and *SNSequate* (Gonzalez, 2014, 2020) implement the Stocking and Lord, Haebara, and moments-based approaches. EQUATE, POLYST, and *plink* can be used with dichotomous and polytomous IRT models (e.g., the GPC, GR, NR models) as well as mixed model calibrations, whereas ST, *SNSequate*, and *equateIRT* are for dichotomous models.

Stocking and Lord (1983) show that if the estimates are error free, then the proper choice of the metric transformation coefficients results in the total characteristic curves from the two forms being linked to coincide. Baker (1996) found that the metric transformation coefficients’ sampling distributions are “well behaved.” As a consequence, one may have confidence in the reasonableness of the metric transformation coefficients; also see Baker (1997). Kaskowitz and de Ayala (2001) studied the impact of error on the total characteristic function method and found the method to be robust to estimation errors. Additional work by Baker and Al-Karni (1991) concluded that the Stocking and Lord total characteristic function procedure is the de facto standard against which other procedures for computing metric transformation coefficients should be compared.

The effect of the metric transformation coefficient κ is to shift the total characteristic function up or down the continuum, whereas the effect of the ζ coefficient is to change the slope of the total characteristic curve. For example, assume that we have two forms whose TCFs are perfectly aligned. If we apply $\kappa = -1$ and $\zeta = 1$ to Form 2’s total characteristic function, then its TCF (the dash line) shifts down the continuum by one logit (Figure 11.2). If we change κ to be 1.5, then Form 2’s total characteristic curve shifts up the scale from its current location by one and one-half logits (Figure 11.3). By changing ζ to 0.5, Form 2’s total characteristic curve becomes steeper (Figure 11.4). Therefore, by proper choice of κ and ζ , we can align Form 2’s total characteristic function with that of Form 1 on our target metric.

The gist of the total characteristic function approach is to determine the values of κ and ζ that results in the forms’ total characteristic functions being minimally different. Stated another way, ζ and κ are chosen to minimize the difference between the forms’ expected trait score estimates by minimizing a loss function. In the ideal case, the expected trait score estimates from two forms for individual i are identical and the loss function is zero. Accordingly, Stocking and Lord (1983) suggest minimizing the loss function⁴

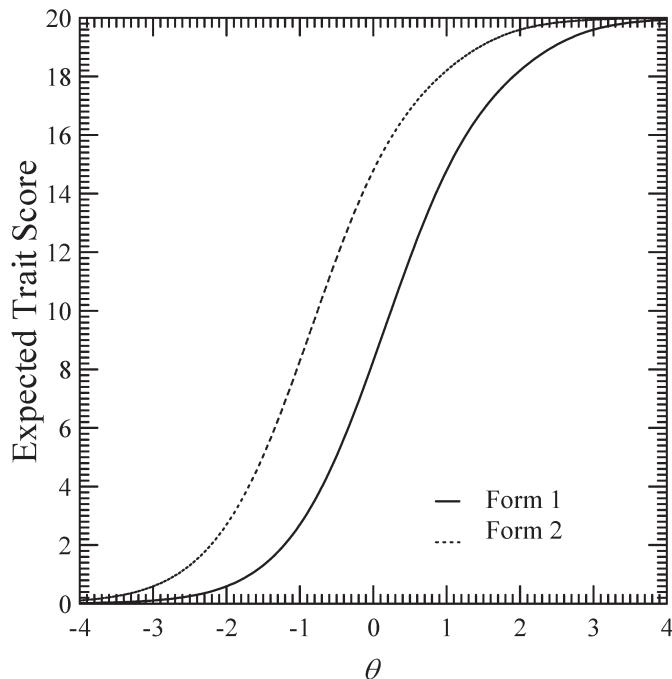


FIGURE 11.2. Total characteristic curves for two forms ($\kappa = -1, \zeta = 1$).

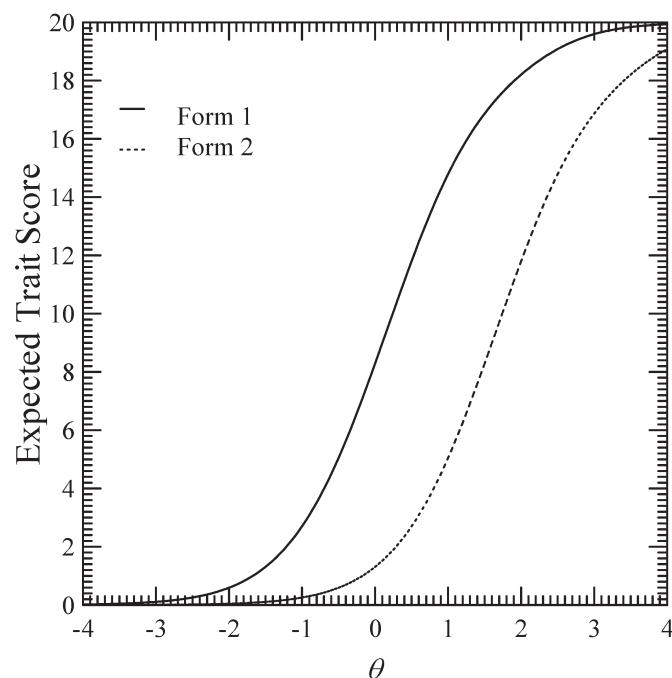


FIGURE 11.3. Total characteristic curves for two forms ($\kappa = 1.5, \zeta = 1$).

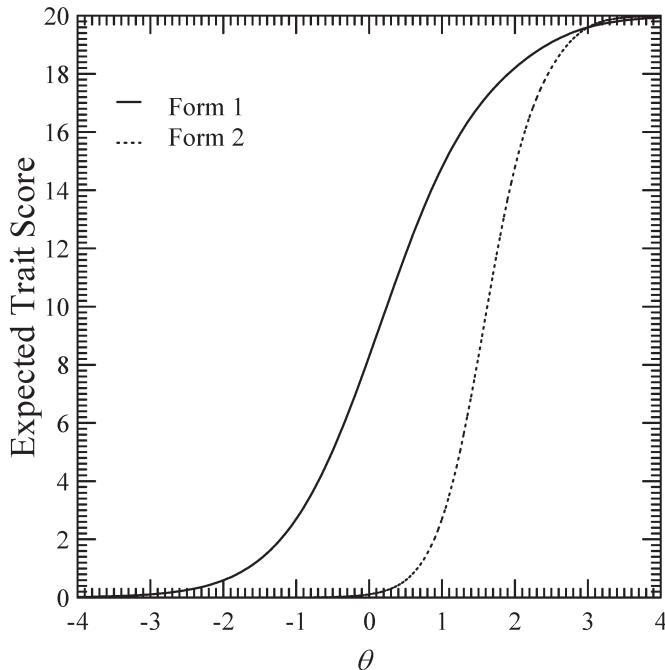


FIGURE 11.4. Total characteristic curves for two forms ($\kappa = 1.5, \zeta = 0.5$).

$$F = \frac{1}{N} \sum_{i=1}^N \left(\hat{T}_{i,1} - \hat{T}_{i,2}^* \right)^2, \quad (11.8)$$

where N is the number of individuals, $\hat{T}_{i,1}$ and $\hat{T}_{i,2}^*$ are the participants' expected trait score estimates on forms 1 and 2, respectively, and

$$\hat{T}_{i,1} = \sum_j p_{ij,1}(\hat{\theta}_{i,1}, \underline{v}_{j,1}) \text{ and } \hat{T}_{i,2}^* = \sum_j p_{ij,2}^2(\hat{\theta}_{i,2}, \underline{v}_{j,2}^*),$$

where the vectors $\underline{v}_{j,2}$ and $\underline{v}_{j,2}^*$ contain common item j 's parameters (or estimates) on Form 1 and the transform Form 2 estimates (i.e., after transformation Form 2 to Form 1's [target] metric given some κ and ζ), respectively. As can be seen, the (quadratic) loss function is the average squared discrepancies between expected trait score estimates (i.e., TCFs). In short, the objective is to determine the values of κ and ζ that when applied to Form 2's item parameters (or estimates) via Equations 11.2 and 11.3 makes individual i 's expected trait score estimates on forms 1 and 2 identical.

To minimize Equation 11.8, we set its derivatives with respect to the transformation coefficients to 0. Following Stocking and Lord (1983) we have the derivatives of F with respect to ζ and κ as

$$\frac{\partial F}{\partial \zeta} = \frac{-2}{N} \sum_{i=1}^N \left(\hat{T}_{i,1} - \hat{T}_{i,2}^* \right) \frac{\partial \hat{T}_{i,2}^*}{\partial \zeta} = 0, \quad (11.9)$$

$$\frac{\partial F}{\partial \kappa} = \frac{-2}{N} \sum_{i=1}^N (\hat{T}_{i,1} - \hat{T}_{i,2}^*) \frac{\partial \hat{T}_{i,2}^*}{\partial \kappa} = 0, \quad (11.10)$$

where

$$\frac{\partial \hat{T}_{i,2}^*}{\partial \zeta} = \sum_{j=1}^L \left(\delta_{j,2} \frac{\partial p_j^*(\theta_i)}{\partial \delta_{j,2}^*} - \frac{\alpha_{j,2}}{\zeta^2} \frac{\partial p_j^*(\theta_i)}{\partial \alpha_{j,2}^*} \right), \quad (11.11)$$

$$\frac{\partial \hat{T}_{i,2}^*}{\partial \kappa} = \sum_{j=1}^L \left(\frac{\partial p_j^*(\theta_i)}{\partial \delta_{j,2}^*} \right), \quad (11.12)$$

$\alpha_{j,2}$ and $\delta_{j,2}$ are item j's unlinked Form 2 discrimination and location parameters, respectively, and p_j^* is appropriate IRT model using the linked item parameter estimates. For example, assume that we are using the 2PL model, then our derivatives (see Lord, 1980)

$$\begin{aligned} \frac{\partial p_j^*(\theta_i)}{\partial \alpha_{j,2}^*} &= (\theta_i - \delta_j^*) p_{ij}^* (1 - p_{ij}^*) \\ \frac{\partial p_j^*(\theta_i)}{\partial \delta_{j,2}^*} &= \alpha_j^* p_{ij}^* (1 - p_{ij}^*) \end{aligned}$$

are substituted into Equations 11.11 and 11.12, which, in turn, are substituted into Equations 11.9 and 11.10.

In contrast to Stocking and Lord's (1983) multivariate search method approach, Baker, Al-Karni, and Al-Dosary (1991) solved for the metric transformation coefficients by using the Davidon–Fletcher–Powell method. To minimize F , $\hat{T}_{i,1}$ and $\hat{T}_{i,2}^*$ are evaluated for a set of N points along the latent variable continuum (e.g., N equidistant theta points between -4 and 4). These N theta points are used in lieu of individuals' θ s to calculate $\hat{T}_{i,1}$ and $\hat{T}_{i,2}^*$. Essentially, this iterative technique begins with provisional metric transformation coefficients ($\kappa^{(1)}$, $\zeta^{(1)}$) that are applied to the initial metric estimates to calculate $\hat{T}_{i,2}^{*(1)}$. The sum of the (squared) discrepancies between $\hat{T}_{i,2}^{*(1)}$ and $\hat{T}_{i,1}$ for each of the N points is determined. In subsequent iterations the previous iteration's metric transformation coefficients ($\kappa^{(t-1)}$, $\zeta^{(t-1)}$) are improved to obtain the t th iteration's $\kappa^{(t)}$ and $\zeta^{(t)}$, F calculated, and convergence assessed. If convergence is not achieved, then another iteration is conducted; otherwise the process terminates. Convergence is achieved when the difference between successive iteration F values are less than the convergence criterion (i.e., comparing $F^{(t)}$ with $F^{(t+1)}$).

One may visualize the result of the minimization of the difference between two total characteristic functions by returning to the TCFs shown in Figures 11.2 through 11.4. For instance, if we apply the metric transformation coefficients $\kappa = -0.005$ and $\zeta = 0.95$ to Form 2's estimates (i.e., through Equations 11.2 and 11.3), then we can transform Form 2's (initial) metric to Form 1's (target) metric. Figure 11.5 shows that after this transformation Form 2's TCF is virtually identical with that of Form 1.

For MIRT models, multidimensional linking may be accomplished in various ways. For example, one might use the total characteristic function method, the item response

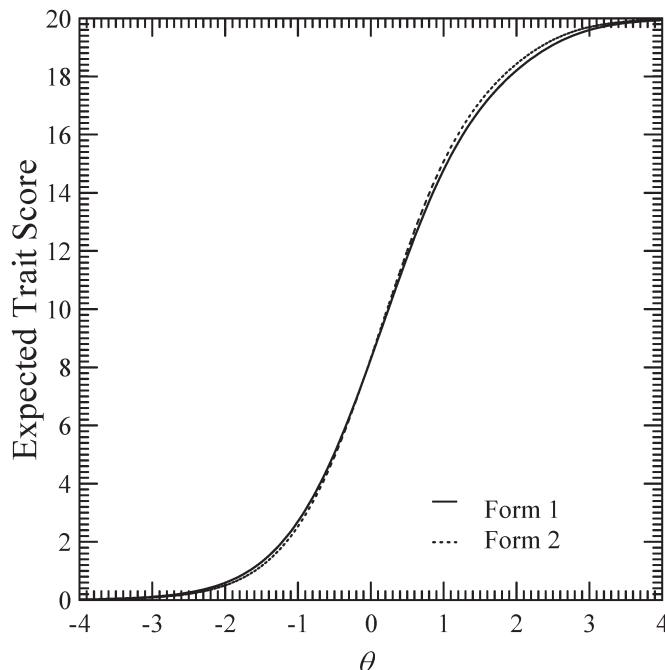


FIGURE 11.5. Total characteristic curves for two forms ($\kappa = -0.005$, $\zeta = 0.95$).

function method, or the equated function method. (The item response function method is analogous to the total characteristic function approach but is focused on minimizing the differences between item response surfaces, and the equated function method is similar to linear equating.) Oshima, Davey, and Lee (2000) studied four different approaches for performing multidimensional linking and found that the total characteristic function and the item response function provided the most stable results, although all four approaches were comparable to one another. They suggested that selecting among the approaches be based on the purpose of performing the linking. For example, if the purpose is to achieve the equivalence of respondents' trait scores regardless of which form is used, then the total characteristic function approach may be the preferred method. However, if one's focus is on differential item functioning (see Chapter 12), then the item response function may be the best choice. (Also see Davey, Oshima, & Lee, 1996, and Hirsch, 1989, for greater detail on multidimensional equating.) Oshima et al.'s (2000) results were for a two-dimensional situation and may not hold for higher dimensional latent spaces.

Example: Application of the Total Characteristic Function Equating Method, EQUATE

In Chapter 4 we provide an example of using the mean-mean approach to linking metrics (see Endnote 14), whereas in Chapter 5 we use the mean-sigma approach (see Table

5.4). Consequently, we now focus on the total characteristic function method for linking and equating. Our data come from the administration of a positivist psychology scale to two groups of 1000 respondents each. Each form is composed of 10 unique items plus 10 items that are in common across forms with all items using a true/false response format. Accordingly, our response data are collected using the common-item nonequivalent groups design. We first use the freeware EQUATE program (Baker, 1993b) to perform the linking and equating of the two metrics. Subsequently, we use SNSequate.

The EQUATE program (Baker, 1993b, 1996) implements the total characteristic function equating technique. (Capitalized “EQUATE” is Baker (1993b, 1996) standalone program, whereas lower case “equate” is Albano’s R package.) The program estimates the ζ and κ coefficients as well as the value of the loss-function F ; in the program ζ is denoted as A, κ is symbolized as K, and these are referred to as METRIC TRANSFORMATION COEFFICIENTS. In addition to calculating the ζ and κ coefficients, the program transforms the item and person parameter estimates from the initial to the target metric. These transformations implement Equations 11.2 and 11.3 using the program’s values for ζ and κ .

In the following, the data associated with the first group is referred to as Form 1 and that of the second group as Form 2. We begin by separately calibrating each form’s data. Our item names reflect to which form an item belongs or if it is a common item. The variable names use the prefix “F1_,” “F2_,” and “C_” for form 1, form 2, and the common items, respectively, plus the item number as a suffix. Every other item beginning with item 2 is a common item. For instance, on Form 1 our first item is F1_1, our second item is the common item C_2, F1_3 is the third item, and so on. In a similar fashion the items on Form 2 are labeled, albeit with the F2_ prefix (i.e., F2_19 is item 19). We use mirt to perform a 2PL model calibration of each form’s data. Although we are not presenting our fit analysis, we evaluated the tenability of IRT’s unidimensionality, conditional independence, and functional form assumptions, as well as performed further model–data fit analysis for each group prior to conducting the equating.

Our R session is presented in Table 11.1. We calibrate Form 1, extract the item parameter estimates, and save them to an ASCII file, form1est.csv, as well as obtain and save our person estimates to an ASCII file, peopleform1.csv. Subsequently, we repeat the process on Form 2 with its item and person parameter estimates saved to form2est.csv and peopleform2.csv, respectively.

Table 11.2 contains the item parameter estimates corresponding to the two administrations. As can be seen, the common items’ estimates for corresponding items vary across forms. In the following we treat the metric defined by Form 1 as our target metric. That is, Form 2’s metric is the initial metric that is transformed to that of Form 1’s (target) metric. Figure 11.6 contains the unlinked TCFs based on the item parameter estimates shown in Table 11.2. As we see, the two curves vary in their slopes and in the locations of their points of inflexion.

EQUATE uses a query-and-answer (Q&A) approach to specify the IRT model, the number of items, the number of common items, and so on. (A command input file could be used in place of this Q&A and piped into the program.) All input files are expected to be in fixed format ASCII (i.e., text) and in the logistic deviate form. Consequently, we

TABLE 11.1. mirt Session for the 2PL Calibration of Positivist Psychology Scale

```

> # load mirt, etc.

===== FORM 1 =====
> form1dat = read.table("Form1.dat", header=TRUE)

> head(form1dat,5)
  case_ID F1_1 C_2 F1_3 C_4 F1_5 C_6 F1_7 ... C_16 F1_17 C_18 F1_19 C_20
1       1   0   0   0   1   0   1   1   ...   1   0   0   0   0   0
2       2   0   1   0   1   0   1   1   ...   1   0   0   0   0   1
3       3   0   0   0   0   0   0   0   ...   0   0   0   0   0   0
4       4   0   1   1   1   0   1   1   ...   1   1   0   0   0   1
5       5   0   0   0   0   0   0   0   ...   0   0   0   0   0   0

> tail(form1dat,5)
  case_ID F1_1 C_2 F1_3 C_4 F1_5 C_6 F1_7 ... C_16 F1_17 C_18 F1_19 C_20
996     996   0   0   0   0   0   0   0   ...   1   0   0   0   0   0
997     997   0   0   0   0   0   0   0   ...   0   0   0   0   0   0
998     998   0   1   0   1   1   0   1   ...   1   0   1   0   0   0
999     999   0   0   0   0   0   0   0   ...   0   0   0   0   0   0
1000    1000   0   0   0   0   0   0   1   ...   0   0   0   0   0   0

> form1dat=within(form1dat,rm(case_ID))                                # remove case label

> print((form1 = mirt(form1dat,1,'2PL')))
  Iteration: 34, Log-Lik: -9230.345, Max-Change: 0.00010

  Call:
  mirt(data = form1dat, model = 1, itemtype = "2PL")

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 34 EM iterations.
  mirt version: 1.31
  M-step optimizer: BFGS
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Log-likelihood = -9230.345
  Estimated parameters: 40
  AIC = 18540.69; AICc = 18544.11
  BIC = 18737; SABIC = 18609.96
  G2 (1048535) = 5672.69, p = 1
  RMSEA = 0, CFI = NaN, TLI = NaN

> print((coef(form1,simplify=TRUE,IRTpars=TRUE)),digits=5)
  $items
      a      b      g      u
F1_1  1.77325  0.91477  0  1
C_2   1.82972 -0.09720  0  1
F1_3  2.73291  0.32710  0  1
C_4   2.12531  0.20129  0  1
F1_5  2.01933  0.81927  0  1
C_6   2.24080 -0.11037  0  1
F1_7  1.82267 -0.33298  0  1
C_8   2.08024  0.46008  0  1
F1_9  1.59371  0.06176  0  1
C_10  2.17044 -0.39067  0  1
F1_11 2.06463 -0.57861  0  1
C_12  2.32767  0.42750  0  1

```

(continued)

TABLE 11.1. (*continued*)

```

F1_13 1.78279 -1.14905 0 1
C_14 1.72176 0.93201 0 1
F1_15 2.27475 -0.21724 0 1
C_16 1.63860 -0.92931 0 1
F1_17 1.49392 0.77336 0 1
C_18 1.55399 1.63373 0 1
F1_19 2.26927 2.60407 0 1
C_20 1.91318 0.54952 0 1

$means
F1
 0

$cov
  F1
F1  1

> # extract Form 1 item parameter estimates. See Table 5.4 or Ch 4, Endnote 14
> # item estimates written to output file:
> write.csv(form1est, file = "form1est.csv")

# obtain person estimates via fscores & display first 6 cases
> head((peopleform1=fscores(form1,method="EAP",full.scores=T,full.scores.SE=T)),6)
      F1      SE_F1
[1,] -0.1715779 0.2657298
[2,]  0.4962998 0.2654102
[3,] -1.8683153 0.5418480
[4,]  0.5014528 0.2656662
[5,] -0.8415580 0.3295935
[6,] -0.6661335 0.3066889

> tail(peopleform1,4)
      F1      SE_F1
[997,] -1.8683153 0.5418480
[998,]  0.6014954 0.2714324
[999,] -1.3907928 0.4291883
[1000,] -1.4363948 0.4391579

> mean(peopleform1[,1])
[1] 7.303037e-05

> sd(peopleform1[,1])
[1] 0.9449186

> # person estimates written to output file:
> write.csv(peopleform1, file = "peopleform1.csv")

=====
> form2dat = read.table("Form2.dat", header=TRUE)

> head(form2dat,5)
  case_ID F2_1 C_2 F2_3 C_4 F2_5 C_6 F2_7 ... C_16 F2_17 C_18 F2_19 C_20
1        1    0    0    0    0    0    0    1 ...    0    0    0    0    0
2        2    0    0    0    0    0    1    1 ...    0    1    0    0    1
3        3    1    0    0    0    0    1    1 ...    1    0    0    0    0
4        4    0    1    0    1    0    0    1 ...    1    0    0    0    0
5        5    0    1    1    1    1    0    1 ...    1    1    1    0    0

```

(continued)

TABLE 11.1. (*continued*)

```

> tail(form2dat,5)
    case_ID F2_1 C_2 F2_3 C_4 F2_5 C_6 F2_7 ... C_16 F2_17 C_18 F2_19 C_20
  996     996   0   0   0   0   0   0   0   ...   0   0   0   0   0   0
  997     997   0   1   1   1   1   1   1   ...   1   0   0   0   0   0
  998     998   0   1   1   1   0   1   0   ...   1   1   0   0   0   1
  999     999   1   1   1   1   0   1   0   ...   1   1   0   0   0   1
 1000    1000   0   0   1   0   0   0   1   ...   1   0   0   0   0   0>

> form2dat=within(form2dat,rm(case_ID)) # remove case label

> print((form2 = mirt(form2dat,1,'2PL')))
  Iteration: 34, Log-Lik: -9840.513, Max-Change: 0.00007

Call:
mirt(data = form2dat, model = 1, itemtype = "2PL")

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 34 EM iterations.
mirt version: 1.31
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Log-likelihood = -9840.513
Estimated parameters: 40
AIC = 19761.03; AICc = 19764.45
BIC = 19957.34; SABIC = 19830.29
G2 (1048535) = 6403.05, p = 1
RMSEA = 0, CFI = NaN, TLI = NaN

> print((coef(form2,simplify=TRUE,IRTpars=TRUE)),digits=5)
$items
      a      b g u
F2_1 1.15876 1.08087 0 1
C_2  1.56580 -0.42157 0 1
F2_3 1.75134 0.16442 0 1
C_4  1.68957 0.05643 0 1
F2_5 1.61408 0.74506 0 1
C_6  1.84599 -0.37063 0 1
F2_7 1.38062 -0.75760 0 1
C_8  1.78499 0.24356 0 1
F2_9 1.39100 -0.27565 0 1
C_10 1.74056 -0.83125 0 1
F2_11 1.77992 -0.91303 0 1
C_12 1.89041 0.30684 0 1
F2_13 1.51689 -1.45088 0 1
C_14 1.31246 0.89401 0 1
F2_15 2.21409 -0.55558 0 1
C_16 1.23835 -1.41666 0 1
F2_17 1.48497 0.57196 0 1
C_18 1.07346 1.87996 0 1
F2_19 2.11805 2.55318 0 1
C_20 1.54353 0.33159 0 1

$means
F1
  0

```

(continued)

TABLE 11.1. (continued)

```

$cov
  F1
F1  1

> # extract Form 2 item parameter estimates. See Table 5.4, 11.6, or Ch 4, Endnote 14
> # item estimates written to output file:
> write.csv(form2est, file = "form2est.csv")

# obtain person estimates via fscores & display first 6 cases
> head((peopleform2$fscores(form2,method="EAP",full.scores=T,full.scores.SE=T)),6)
      F1      SE_F1
[1,] -0.602118044 0.3222754
[2,] -0.296028265 0.3126320
[3,]  0.003882136 0.3132450
[4,] -0.356900940 0.3136942
[5,]  0.258831123 0.3211096
[6,]  1.130579305 0.3934367

> tail(peopleform2,4)
      F1      SE_F1
[997,]  0.5100066 0.3352582
[998,]  0.3819403 0.3272633
[999,]  0.7237828 0.3520379
[1000,] -0.4786136 0.3170756

> mean(peopleform2[,1])
[1] 1.986856e-06

> sd(peopleform2[,1])
[1] 0.9331412

> # person estimates written to output file:
> write.csv(peopleform2, file = "peopleform2.csv")

```

convert `form1est.csv` and `form1est.csv` to be in fixed format by aligning all values column-wise. Form 1's estimates are saved to a new file called `Form1estfxdfmt.dat`, whereas Form 2's estimates are in `Form2estfxdfmt.dat`.

If the program is used to simply link two metrics, then one has only two input files containing the item parameter estimates on the target and initial metrics. In EQUATE's parlance, Form 2's (initial) metric is the "FROM" metric and Form 1's target metric is the "TO" metric. However, because we are using EQUATE for both linking and equating, we have a third file containing the Form 2 person location estimates to be used in equating. Specifically, one of our item parameter estimate files is called `Form1estfxdfmt.dat` and contains the item parameter estimates that define the target metric (i.e., the "TO" metric). Our other item parameter estimate file, `Form2estfxdfmt.dat`, contains the item parameter estimates that are on the initial metric (i.e., the "FROM" metric). The third fixed format file that we need, `Peopleform2fxdfmt.dat`, contains the person location estimates from the Form 2 administration and, as a result, are on the initial metric.

Table 11.3 contains the Q&A session for specifying the analysis. As can be seen,

we start with the analysis's title and specify that EQUATE should use 50 points in its minimization of Equation 11.8 (i.e., ENTER NUMBER OF ABILITY SCALE POINTS N=). The subsequent queries allow us to specify the IRT model, the number of items, and the FORTRAN format for the initial metric. (FORTRAN formats are briefly discussed in Appendix G, "FORTRAN Formats.") Following the initial metric information, we provide analogous information for the target metric. We instruct the program to create a file (Form2star.dat) to contain the linked Form 2 estimates. Because we want EQUATE to also perform equating, we answer "yes" (i.e., "Y") to the TRANSFORM THETAS? Y/N prompt. As a result, we identify the file containing the person location estimates on the initial metric, Peopleform2fxdfmt.dat, as well as specify a file, Peopleform2star.dat, for the equated person location estimates. At the end of the Q&A we are prompted to provide the positions of the common items on the initial (i.e.,

TABLE 11.2. Item Parameter Estimates for Two Forms

Form 1 (target metric)		Form 2 (initial metric)			
item	$\hat{\alpha}_{i,1}$	$\hat{\delta}_{i,1}$	item	$\hat{\alpha}_{i,2}$	$\hat{\delta}_{i,2}$
F1_1	1.7733	0.9148	F2_1	1.1588	1.0809
C_2	1.8297	-0.0972	C_2	1.5658	-0.4216
F1_3	2.7329	0.3271	F2_3	1.7513	0.1644
C_4	2.1253	0.2013	C_4	1.6896	0.0564
F1_5	2.0193	0.8193	F2_5	1.6141	0.7451
C_6	2.2408	-0.1104	C_6	1.8460	-0.3706
F1_7	1.8227	-0.3330	F2_7	1.3806	-0.7576
C_8	2.0802	0.4601	C_8	1.7850	0.2436
F1_9	1.5937	0.0618	F2_9	1.3910	-0.2757
C_10	2.1704	-0.3907	C_10	1.7406	-0.8313
F1_11	2.0646	-0.5786	F2_11	1.7799	-0.9130
C_12	2.3277	0.4275	C_12	1.8904	0.3068
F1_13	1.7828	-1.1491	F2_13	1.5169	-1.4509
C_14	1.7218	0.9320	C_14	1.3125	0.8940
F1_15	2.2748	-0.2172	F2_15	2.2141	-0.5556
C_16	1.6386	-0.9293	C_16	1.2384	-1.4167
F1_17	1.4939	0.7734	F2_17	1.4850	0.5720
C_18	1.5540	1.6337	C_18	1.0735	1.8800
F1_19	2.2693	2.6041	F2_19	2.1181	2.5532
C_20	1.9132	0.5495	C_20	1.5435	0.3316

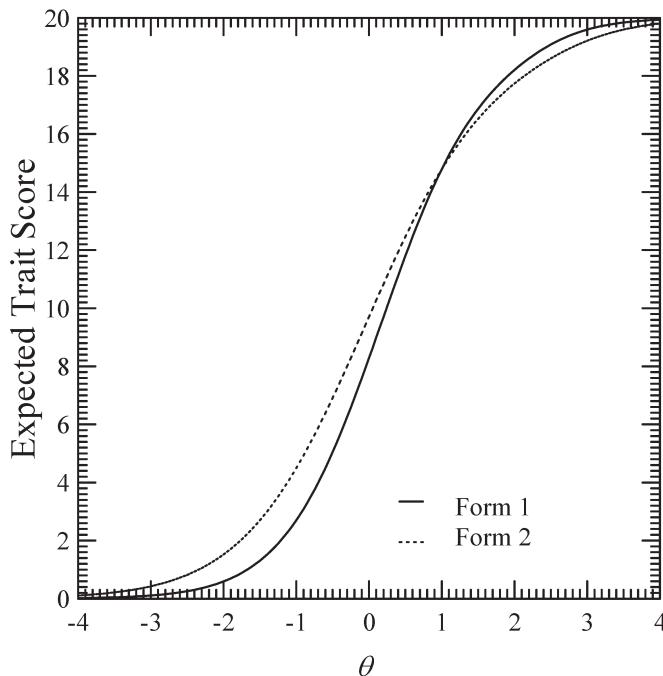


FIGURE 11.6. Unequated TCFs for the two forms shown in Table 11.2.

SPECIFY ANCHOR ITEM IDs FOR THE “FROM” INSTRUMENT and target (i.e., SPECIFY ANCHOR ITEM IDs FOR THE “TO” METRIC) metrics. In this example, even numbered items are the common items and have the same position on both forms. However, to be clear, the common items approach does not require the common items be in the same position on both forms.

After the Q&A session, EQUATE performs its analysis. The abridged output is presented in Table 11.4 and the linked item parameter estimates and equated person locations estimates are shown in Table 11.5. EQUATE echoes the input specifications and its initial (start) values for κ (K) and ζ (A). After iterating three times, the program provides its estimates for the metric transformation coefficients, $\kappa = 0.2235$ and $\zeta = 0.7987$. Descriptive statistics for our Form 2 item parameter estimates, after being placed on the target metric, are provided, along with similar statistics for the equated person location estimates.

The left panel of Table 11.5 shows the results of Form 2’s item parameter estimates after they have been linked to the target (Form 1’s) metric; this is the contents of the file Form2star.dat. These values are obtained by using $\kappa = 0.2235$ and $\zeta = 0.7987$ in Equations 11.2 and 11.3, respectively, with the item parameter estimates listed in the right panel of Table 11.2. Table 11.5’s right panel contains the unequated ($\hat{\theta}$) and the equated person location estimates ($\hat{\theta}^*$) for the first 20 cases of Peopleform2star.dat. Using the linked Form 2 item parameter estimates (i.e., $\hat{\alpha}^*$ and $\hat{\delta}^*$) and Form 1’s item parameter estimates to calculate \hat{T}_{i1} and $\hat{T}_{i,2}^*$ we see that the two corresponding TCFs are indistinguishable (Figure 11.7).

TABLE 11.3. EQUATE's Query-and-Answer Session^a

TYPE A TITLE FOR THE COMPUTER RUN Example TCF equating, 2PL, 10 common items	← input title
ENTER NUMBER OF ABILITY SCALE POINTS N= 50	← number of points used in the minimization
RESPONSE MODE DICHOTOMOUS, GRADED OR NOMINAL? D/G/N D	← dichotomous model
ENTER NUMBER OF PARAMETERS IN ICC MODEL 1, 2, 3: 2	← model: 2 parameter
ENTER NAME OF "FROM METRIC" ITEM PARAMETER FILE Form2estfmt.dat	← the initial metric (untransformed)
ENTER FORMAT OF "FROM METRIC FILE" (3X,F6.4,1X,F7.4)	← FORTRAN format: $\hat{\alpha}_i$ read as F6.4 & $\hat{\delta}_i$ read as F7.4
ENTER NUMBER OF ITEMS IN "FROM" TEST 20	
IS FROM METRIC LOGISTIC OR NORMAL? L/N L	← with 2 parameters entered above model: 2PL
ENTER NAME OF "TO METRIC" ITEM PARAMETER FILE Formtestfmt.dat	← the target metric
ENTER FORMAT OF "TO METRIC FILE" (3X,F6.4,1X,F7.4)	← FORTRAN format: $\hat{\alpha}_i$ read as F6.4 & $\hat{\delta}_i$ read as F7.4
ENTER NUMBER OF ITEMS IN "TO" TEST 20	
IS TO METRIC LOGISTIC OR NORMAL? L/N L	
ENTER NAME OF FILE TO STORE TRANSFORMED ITEM PARAMETERS Form2star.dat	← the file containing Form 2's linked (transformed) item parameter estimates
TRANSFORM THETAS? Y/N Y	← perform equating
ENTER NAME OF "FROM" THETA FILE Peopleform2fxdfmt.Dat	← file containing the (untransformed) $\hat{\theta}$ s on initial metric (i.e., person who took Form 2)
ENTER FORMAT OF "FROM" THETA FILE (F7.4)	← FORTRAN format to read $\hat{\theta}$ s
ENTER NUMBER OF EXAMINEES 40	← number of individuals to be equated to Form 1's metric
ENTER NAME OF FILE FOR TRANSFORMED THETAS Peopleform2star.dat	← file containing the equated (transformed) $\hat{\theta}^*$ s
ARE THESE SPECIFICATIONS OK? Y/N Y	
SPECIFY ANCHOR ITEM IDs FOR THE "FROM" INSTRUMENT ENTER LIST OF ANCHOR ITEMS SEPARATE WITH COMMAS TERMINATE WITH COLON 2,4,6,8,10,12,14,16,18,20:	← ordinal positions of common items.
SPECIFY ANCHOR ITEM IDs FOR THE "TO" METRIC ENTER LIST OF ANCHOR ITEMS SEPARATE WITH COMMAS TERMINATE WITH COLON 2,4,6,8,10,12,14,16,18,20:	← ordinal positions of common items.

^aThe text following the '←' is provided to help the reader understand the corresponding input.

TABLE 11.4. Abridged EQUATE Output^a

```

Example TCF equating, 2PL, 10 common items
NUMBER OF ABILITY SCALE POINTS=      50
DICHOTOMOUS RESPONSE MODEL
ICC MODEL HAS      2 PARAMETERS

:
                                         ← Echoing of input specifications

INITIAL VALUE FOR A=      0.8002   INITIAL VALUE FOR K=      0.2139
FUNCTION AT INITIAL VALUES =      0.000990
NUMBER OF ITERATIONS PERFORMED =      3
METRIC TRANSFORMATION COEFFICIENTS ARE
A=      0.7987   K=      0.2235 ← The metric transformation coefficients: A is  $\zeta$  & K is  $\kappa$ 
FUNCTION VALUE =      0.000692

SUMMARY STATISTICS FOR TRANSFORMED ITEMS
MEAN B=      0.297  VARIANCE B=      0.669  STD DEV B=      0.818
MEAN A=      2.009  VARIANCE A=      0.141  STD DEV A=      0.375

SUMMARY STATISTICS OF TRANSFORMED ABILITIES
MEAN =      0.223477
VARIANCE=      0.555494
STANDARD DEVIATION=      0.745315

:

```

^aThe text following the '←' is provided to help the reader understand the corresponding input.

Example: Application of the Total Characteristic Function Equating Method, SNSequate

The R package SNSequate (Gonzalez, 2014, 2020) implements traditional equating methods (e.g., mean, linear, equipercentile equating methods) and IRT-based approaches (e.g., mean-mean, mean-sigma, Haebara, Stocking-Lord TCF), as well as kernel equating using various kernel methods (see Appendix G). Additionally, it implements local equating and asymmetric item characteristic functions.

Our SNSequate R session is shown in Table 11.6. SNSequate requires the item parameter estimates conform to the three-parameter model logistic deviate form and the forms to be linked to reside in the same input object. We address the first requirement by first extracting each form's item parameter estimates in the logistic deviate format (`coef(form1,simplify = TRUE, IRTpars = TRUE)$items[,c('a','b')]`).⁵ Second, we augment each form's 2PL model estimates by a null vector representing χ that contains only zeroes. We create this vector using the `replicate` function (`cc = rep(0,20)`) and merge it with each form's item parameter estimate file using the `cbind` function (`cbind(form1est,cc = rep(0,20))`). To create a single-input object, `forms`, we merge our augmented item parameter estimate files by `cbind(form1est,form2est)`.

TABLE 11.5. Linked Form 2 Item Parameter Estimates and Equated Person Location Estimates

Form 2 (linked)		Unequated	Equated
item	$\hat{\alpha}_{i,2}^*$	$\hat{\delta}_{i,2}^*$	$\hat{\theta}$
1	1.451	1.0872	-0.6021
2	1.960	-0.113	-0.2960
3	2.193	0.355	0.0039
4	2.115	0.269	-0.3569
5	2.021	0.819	0.2588
6	2.311	-0.073	1.1306
7	1.729	-0.382	0.3374
8	2.235	0.418	-1.1563
9	1.742	0.003	-0.5055
10	2.179	-0.440	-0.2630
11	2.228	-0.506	1.3632
12	2.367	0.469	0.0259
13	1.899	-0.935	0.2369
14	1.643	0.938	-0.1989
15	2.772	-0.220	0.0100
16	1.550	-0.908	-0.2957
17	1.859	0.680	0.6677
18	1.344	1.725	-0.0414
19	2.652	2.263	0.1735
20	1.932	0.488	-0.8222
			-0.433

As was the case with EQUATE, we need to specify the common items for SNSequate. Because our common items are the even-number items on each form, we use the sequence function to create the even-number series between our first common item C _ 2 and our last one C _ 20 (seq(2,20,2)) and store the series in anchor.

In our call to the `irt.link` function, we pass our `forms` object, the anchor vector, and specify the `model` to be 2PL and the logistic metric. As can be seen, `irt.link` automatically produces the metric coefficients based on the mean-mean, mean-sigma, Haebara, as well as the Stocking and Lord approaches. Our interest is in the Stocking-Lord line that shows that $\zeta(A)$ is 0.7996 and $\kappa(K)$ equals 0.22201. To link Form 2's metric to that of Form 1, we extract and assign these values to `zeta` and `kappa`,

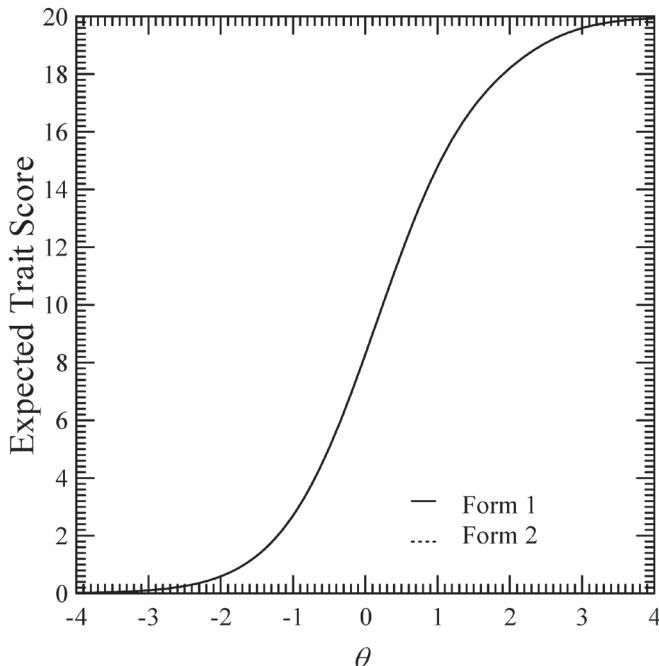


FIGURE 11.7. TCCs after initial metric (Form 2) is linked to the target metric (Form 1).

respectively. We first transform our Form 2 item location estimates using Equation 11.2 (`form2bstar = form2est$b*zeta+kappa`), followed by using Equation 11.3 to transform Form 2's item discrimination estimates (`form2astar = form2est$a/zeta`). To equate Form 2's person location estimates (`peopleform2`) with Form 1, we apply Equation 11.1 with ξ representing Form 2's $\hat{\theta}$ s (i.e., `peopleform2[,1]*zeta + kappa`). Comparing our equated Form 2 $\hat{\theta}^*$ s with the unequated $\hat{\theta}$ s, we see, for example, that the first person who took Form 2 was originally estimated to be located at -0.602 (see Table 11.1), but once we take into account the differences in the forms' metrics, we estimate the person to be higher on the positive psychology scale ($\hat{\theta}_1^* = -0.259$) than before equating.

Example: Fixed-Item and Concurrent Calibration Equating

For pedagogical reasons we demonstrate the fixed-item parameter approach followed by concurrent (simultaneous) calibration. In our dichotomous case, `mirt` estimates the slope–intercept format of the four-parameter model (cf. Chalmers, 2017). Accordingly, we need to fix our common items using their slope–intercept format. For convenience we present Form 1's item parameter estimates in their slope–intercept format in Table 11.7. Our interest is only in the common items (i.e., C _ 2, C _ 4, etc.). To apply the fixed-item parameter method to Form 2's estimate, we create a model specification object in which all items load on our positive psychology construct ($F = 1\text{--}20$), specify the

TABLE 11.6. SNSequate Session for Linking and Equating Example

```

> # This is a continuation of the session from Table 11.1

> library(SNSequate)
  Loading required package: magic
  Loading required package: abind
  Warning message:
  package 'SNSequate' was built under R version 3.6.3

> # extract 2PL model item parameter estimates, logistic deviate format
> form1est=coef(form1,simplify=TRUE,IRTpars=TRUE)$items[,c('a','b')]
> form1est=as.data.frame(form1est)

> form2est=coef(form2,simplify=TRUE,IRTpars=TRUE)$items[,c('a','b')]
> form2est=as.data.frame(form2est)

> # add pseudo-guessing parameter to each form with value set to 0
> form1est=cbind(form1est,cc =rep(0,20)) ; form2est=cbind(form2est,cc =rep(0,20))

> form1est
      a          b cc
F1_1  1.773247  0.91476834  0
C_2   1.829720 -0.09720274  0
F1_3   2.732906  0.32709646  0
C_4   2.125312  0.20128634  0
:       :          :      :
F1_19  2.269270  2.60407195  0
C_20  1.913181  0.54952307  0

> form2est
      a          b cc
F2_1  1.158759  1.08086684  0
C_2   1.565799 -0.42157483  0
F2_3   1.751338  0.16442202  0
C_4   1.689571  0.05643191  0
:       :          :      :
F2_19  2.118051  2.55317553  0
C_20  1.543531  0.33159176  0

> forms=cbind(form1est,form2est) # merge forms 1 and 2
> forms
      a          b cc      a          b cc
 1  1.773247  0.91476834  0  1.158759  1.08086684  0
 2  1.829720 -0.09720274  0  1.565799 -0.42157483  0
 3  2.732906  0.32709646  0  1.751338  0.16442202  0
 4  2.125312  0.20128634  0  1.689571  0.05643191  0
:       :          :      :       :          :
19  2.269270  2.60407195  0  2.118051  2.55317553  0
20  1.913181  0.54952307  0  1.543531  0.33159176  0

> anchor=seq(2,20,2) # create vector containing common items; items 2-20 in steps
of 2
> anchor
[1]  2  4  6  8 10 12 14 16 18 20

> # link forms

```

(continued)

TABLE 11.6. (continued)

```

> print((lnksnl=irt.link(forms, anchor, model = "2PL",icc = "logistic", D = 1.7)))

Call:
  irt.link.default(parm = forms, common = anchor, model = "2PL",
    icc = "logistic", D = 1.7)

IRT parameter-linking constants:
      A           B
Mean-Mean   0.8001913 0.2138638 # A is  $\zeta$  & B is  $\kappa$ 
Mean-Sigma   0.7793053 0.2152679
Haebara     0.7812961 0.2161181
Stocking-Lord 0.7995614 0.2220930

> lnksnl$StockLord
[1] 0.7995614 0.2220930

> zeta=lnksnl$StockLord[1]

> kappa=lnksnl$StockLord[2]

> # application of Eq 11.2 to transform Form 2 item location estimates
> form2bstar=form2est$b*zeta+kappa

> # application of Eq 11.3 to transform Form 2 item discrimination estimates
> form2astar=form2est$a/zeta

> form2star=cbind( form2astar,form2bstar) # create linked Form 2 estimates
> form2star
      form2astar  form2bstar
 [1,]  1.449244  1.08631244
 [2,]  1.958322 -0.11498191
 [3,]  2.190374  0.35355854
 [4,]          :
 [5,]          :
 [6,]          :
 [7,]          :
 [8,]          :
 [9,]          :
 [10,]         :
 [11,]         :
 [12,]         :
 [13,]         :
 [14,]         :
 [15,]         :
 [16,]         :
 [17,]         :
 [18,]         :
 [19,]         :
 [20,]  1.930472  0.48722102

> # Equate form 2 to form 1
> # application of Eq 11.1 with casi=theta hat; equate Form 2 theta hats
  to Form 1's
> head((t_est_star=as.data.frame(peopleform2[,1])*zeta+kappa),6)
  peopleform2[, 1]
  1       -0.25933730
  2       -0.01459973
  3        0.22519705
  4       -0.06327117
  5        0.42904442
  6       1.12606061

> tail(t_est_star,4)
  peopleform2[, 1]
  997       0.6298747
  998       0.5274777
  999       0.8008018
 1000      -0.1605879

> # item estimates written to output file:
> write.csv(form2star, file = "form2eststar.csv")

> # personestimates written to output file:
> write.csv(t_est_star, file = "peopleform2star.csv")

```

starting values for each common item's discrimination and intercept, and then hold the slope and intercept fixed at the starting values. For example, for the first common item C_2, its Form 1 estimates are $\hat{\alpha}_2 = 1.82972$ and $\hat{\gamma}_2 = 0.17785$. Therefore, we provide these estimates as the starting values (START = (C_2,a1,1.82972), (C_2,d,0.17785)) and subsequently hold them fixed at these values (FIXED = (C_2,a1), (C_2,d)); the population mean and covariance are estimated. (In mirt nomenclature, a1 is our alpha and d is our gamma.) For our Form 2 calibration we pass our model specification to mirt.

The resulting Form 2 item parameter estimates (form2fxdest) are on Form 1's metric. As a check that our common items were held fixed, we compare Form 2 common item estimates with their values on Form 1 (see Table 11.1). For instance, Form 2's C_20 of $\hat{\alpha}_2 = 1.91318$ and $\hat{\delta}_2 = 0.54952$ match those obtained on Form 1; a similar result would be obtained if we compared corresponding common item $\hat{\gamma}_j$ s. All Form 2 person location estimates are on Form 1's metric.

For our concurrent calibration example, we need to set up our calibration data following a simple construction principle: all items in common across forms are located in the same response column, and all noncommon items occupy unique response columns. Figure 11.1 demonstrates one way of implementing this principle. One can simultaneously calibrate more than two response data sets by following our construction principle.

For convenience of presentation, we move all of Form 1's common items to follow the noncommon items, whereas for Form 2 we do the opposite. Consequently, on Form 1 our items are F1_1, F1_3, F1_5, ..., F1_19, C_2, C_4, ..., C_20, and for Form 2 we have C_2, C_4, ..., C_20, F2_1, F2_3, F2_5, ..., F2_19. Upon concatenating our response data files our file has a layout that corresponds to the schematic in Figure 11.1. (Of course, we also could have not moved any Form 1 items and simply moved Form 2's common items below the corresponding Form 1 common item and left Form 2's unique items in their original positions.) Our calibration data file contains 2000 cases and 30 items, 10 of which are common items. Because we calibrate our data using an R package, we address the missing data created by not presenting any item to an individual by using NA for each missing response.

Table 11.8 contains our R session. (Although not conducted here, in practice we would assess the tenability of our model's assumptions [e.g., unidimensionality] and conduct both model- and item-level fit analyses before accepting our item parameter estimates.) After removing our case_ID variable, we call mirt specifying a 2PL model calibration of the response data forms1_2. Our estimates for all 30 items as well as our equated person estimates follow. Unlike the TCF and fixed-item parameter equating methods, our common items C_2, C_4, ..., C_20 are estimated on the basis of all 2000 cases, whereas the non-common items are estimated using only those individuals who responded to a form (i.e., 1000). However, as is the case with the above methods, our person $\hat{\theta}$ s are estimated using the common items and the unique items on the form they took. Of course, all 2000 respondents are on the same metric.

Examination of the item and person estimates across the different analysis shows that the approaches that used Form 1 as the target metric (i.e., EQUATE, SNSequate, and the fixed-item parameter) have Form 2 $\hat{\alpha}_j^*$ s and $\hat{\delta}_j^*$ s that are essentially equivalent.

TABLE 11.7. Fixed-Item Parameter Session for Linking and Equating Example

```

> # This is a continuation of the session from Table 11.1

> print((coef(form1,simplify=TRUE)),digits=5)                                # slope-intercept format
$items
    a1      d g u
F1_1  1.77325 -1.62211 0 1
C_2   1.82972  0.17785 0 1
F1_3   2.73291 -0.89392 0 1
C_4   2.12531 -0.42780 0 1
F1_5   2.01933 -1.65438 0 1
C_6   2.24080  0.24732 0 1
F1_7   1.82267  0.60691 0 1
C_8   2.08024 -0.95708 0 1
F1_9   1.59371 -0.09843 0 1
C_10  2.17044  0.84792 0 1
F1_11  2.06463  1.19462 0 1
C_12  2.32767 -0.99509 0 1
F1_13  1.78279  2.04852 0 1
C_14  1.72176 -1.60470 0 1
F1_15  2.27475  0.49417 0 1
C_16  1.63860  1.52276 0 1
F1_17  1.49392 -1.15533 0 1
C_18  1.55399 -2.53881 0 1
F1_19  2.26927 -5.90934 0 1
C_20  1.91318 -1.05134 0 1

$means
F1
  0

$cov
  F1
F1  1

> modspec2='F = 1-20
+ START=(C_2,a1,1.82972), (C_2,d, 0.17785)
+ FIXED=(C_2,a1), (C_2,d)
+ START=(C_4,a1,2.12531), (C_4,d, -0.42780)
+ FIXED=(C_4,a1), (C_4,d)
+ START=(C_6,a1,2.24080), (C_6,d, 0.24732)
+ FIXED=(C_6,a1), (C_6,d)
+ START=(C_8,a1,2.08024), (C_8,d, -0.95708)
+ FIXED=(C_8,a1), (C_8,d)
+ START=(C_10,a1,2.17044), (C_10,d, 0.84792)
+ FIXED=(C_10,a1), (C_10,d)
+ START=(C_12,a1,2.32767), (C_12,d,-0.99509)
+ FIXED=(C_12,a1), (C_12,d)
+ START=(C_14,a1,1.72176), (C_14,d,-1.60470)
+ FIXED=(C_14,a1), (C_14,d)
+ START=(C_16,a1,1.63860), (C_16,d, 1.52276)
+ FIXED=(C_16,a1), (C_16,d)
+ START=(C_18,a1,1.55399), (C_18,d,-2.53881)
+ FIXED=(C_18,a1), (C_18,d)
+ START=(C_20,a1,1.91318), (C_20,d,-1.05134)
+ FIXED=(C_20,a1), (C_20,d)
+ FREE = (GROUP, MEAN_1)
+ FREE = (GROUP, COV_11)'


```

(continued)

TABLE 11.7. (*continued*)

```

> print((form2fxd = mirt(form2dat, model=modspec2, '2PL')))
  Iteration: 19, Log-Lik: -9849.038, Max-Change: 0.00010

  Call:
  mirt(data = form2dat, model = modspec2, itemtype = "2PL")

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 19 EM iterations.
  mirt version: 1.30
  M-step optimizer: nlmminb
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Log-likelihood = -9849.038
  Estimated parameters: 22
  AIC = 19742.08; AICc = 19743.11
  BIC = 19850.05; SABIC = 19780.17
  G2 (1048553) = 6420.1, p = 1
  RMSEA = 0, CFI = NaN, TLI = NaN
>

> Form 2 estimates on Form 1's metric
> print((form2fxdest=coef(form2fxd,simplify=TRUE,IRTpars=TRUE)),digits=5)
  $items
    a          b g u
  F2_1  1.43937  1.09383 0 1
  C_2   1.82972 -0.09720 0 1
  F2_3   2.16471  0.35623 0 1
  C_4   2.12531  0.20129 0 1
  F2_5   2.00510  0.82372 0 1
  C_6   2.24080 -0.11037 0 1
  F2_7   1.71317 -0.38736 0 1
  C_8   2.08024  0.46008 0 1
  F2_9   1.72549  0.00112 0 1
  C_10  2.17044 -0.39067 0 1
  F2_11  2.20877 -0.51277 0 1
  C_12  2.32767  0.42750 0 1
  F2_13  1.88350 -0.94570 0 1
  C_14  1.72176  0.93201 0 1
  F2_15  2.74326 -0.22500 0 1
  C_16  1.63860 -0.92931 0 1
  F2_17  1.84102  0.68470 0 1
  C_18  1.55399  1.63374 0 1
  F2_19  2.65464  2.26882 0 1
  C_20  1.91318  0.54952 0 1

  $means
    F
  0.22316

  $cov
    F
  F  0.64847

```

(continued)

TABLE 11.7. (continued)

```
# obtain person estimates via fscores & display first 6 cases
> head((peopleform2fxd=fscores(form2fxd,method="EAP",full.scores=T,
+ full.scores.SE=T)),6)
      F1      SE_F1
[1,] -0.26357139 0.2602740
[2,] -0.01365576 0.2527620
[3,]  0.23027426 0.2535260
[4,] -0.07404458 0.2537726
[5,]  0.45011629 0.2604701
[6,]  1.13225227 0.3145896

> tail(peopleform2fxd,4)
      F1      SE_F1
[997,]  0.6430055 0.2711831
[998,]  0.5307381 0.2644359
[999,]  0.7953131 0.2824935
[1000,] -0.1574410 0.2559979

> mean(peopleform2fxd[,1])
[1] 0.2231649

> sd(peopleform2fxd[,1])
[1] 0.7515294

> # person estimates written to output file:
> write.csv(peopleform2fxd, file = "peopleform2fxd.csv")
```

However, when the comparison involves the concurrent calibration estimates, there appears to be less agreement. This is because the concurrent calibration method produces its own metric. Consequently, we cannot directly compare its estimates to those of the TCF and fixed-item parameter equating methods without transforming to a target metric. However, we can compare the estimates in terms of their linearity. The correlations between the concurrent calibration's $\hat{\theta}$ s for the second form and the equated $\hat{\theta}$ s from EQUATE, SNSequate, and the fixed-item parameter approach are each 0.9998 with corresponding scatterplots that show highly linear relationships. With respect to item parameter estimates, the correlations between the concurrent calibration's $\hat{\alpha}_j$ s and $\hat{\delta}_j$ s with EQUATE's $\hat{\alpha}_j^*$ s and $\hat{\delta}_j^*$ s are 0.981 and 0.999, respectively, and with the fixed-item parameter method they are 0.977 for $\hat{\alpha}_j$ s and 0.999 for $\hat{\delta}_j$ s. For completeness, the correlations between EQUATE's $\hat{\alpha}_j^*$ s and $\hat{\delta}_j^*$ s with those of the fixed-item parameter approach are 0.982 and 0.999, respectively. As might be expected given the close agreement between EQUATE's and SNSequate's ζ and κ values, EQUATE's $\hat{\alpha}_j^*$ s and $\hat{\delta}_j^*$ s correlate perfectly with those of SNSequate.

Summary

Whenever we use different forms of an instrument and/or different groups of people, then our estimates will potentially and most likely be on different metrics. If we need to align the different item parameter estimate metrics with one another, then the process

TABLE 11.8. Concurrent Calibration Session for Linking and Equating Example

```

> forms1_2 = read.table("Forms1_2NA.dat", header=TRUE)

> head(forms1_2,5)
  case_ID F1_1 F1_3 F1_5 ... F1_19 C_2 C_4 ... C_20 F2_1 F2_3 F2_5 ... F2_19
1       1     0     0     0      0     0     1      0    NA    NA    NA      NA
2       2     0     0     0      0     1     1      1    NA    NA    NA      NA
3       3     0     0     0      0     0     0      0    NA    NA    NA      NA
4       4     0     1     0      0     1     1      1    NA    NA    NA      NA
5       5     0     0     0      0     0     0      0    NA    NA    NA      NA

> tail(forms1_2,5)
  case_ID F1_1 F1_3 F1_5 ... F1_19 C_2 C_4 ... C_20 F2_1 F2_3 F2_5 ... F2_19
1996    1996   NA   NA   NA      NA     0     0      0    0    0    0      0
1997    1997   NA   NA   NA      NA     1     1      0    0    1    1      0
1998    1998   NA   NA   NA      NA     1     1      1    0    1    0      0
1999    1999   NA   NA   NA      NA     1     1      1    1    1    0      0
2000    2000   NA   NA   NA      NA     0     0      0    0    1    0      0

> forms1_2=within(forms1_2,rm(case_ID)) # remove case label

> concurrent = mirt(forms1_2,1,'2PL',SE=T,SE.type='Fisher')
  Iteration: 48, Log-Lik: -19097.688, Max-Change: 0.00009

  Calculating information matrix...
  Error: cannot allocate vector of size 4.0 Gb
  >

> print((concurrent = mirt(forms1_2,1,'2PL')))
  Iteration: 48, Log-Lik: -19097.688, Max-Change: 0.00009

  Call:
  mirt(data = forms1_2, model = 1, itemtype = "2PL")

  Full-information item factor analysis with 1 factor(s).
  Converged within 1e-04 tolerance after 48 EM iterations.
  mirt version: 1.31
  M-step optimizer: BFGS
  EM acceleration: Ramsay
  Number of rectangular quadrature: 61
  Latent density type: Gaussian

  Log-likelihood = -19097.69
  Estimated parameters: 60
  AIC = 38315.38; AICc = 38319.15
  BIC = 38651.43; SABIC = 38460.81

> print((forms1_2est=coef(concurrent,simplify=TRUE,IRTpars=TRUE)),digits=5)
  $items
    a      b g u
  F1_1  1.68062  0.86362 0 1
  F1_3  2.59350  0.24188 0 1
  F1_5  1.91138  0.76321 0 1
  F1_7  1.74657 -0.45098 0 1
  F1_9  1.51776 -0.03662 0 1
  F1_11 1.97823 -0.70848 0 1
  F1_13 1.71845 -1.29787 0 1

```

(continued)

TABLE 11.8. (*continued*)

```

F1_15 2.16939 -0.33143 0 1
F1_17 1.41970 0.71320 0 1
F1_19 2.17940 2.62724 0 1
C_2 1.71877 -0.24805 0 1
C_4 1.90026 0.13389 0 1
C_6 2.05406 -0.23107 0 1
C_8 1.94630 0.35342 0 1
C_10 1.97787 -0.58681 0 1
C_12 2.10332 0.36986 0 1
C_14 1.51788 0.90744 0 1
C_16 1.45132 -1.13402 0 1
C_18 1.30843 1.72436 0 1
C_20 1.73872 0.44232 0 1
F2_1 1.23190 1.11345 0 1
F2_3 1.86142 0.25301 0 1
F2_5 1.71839 0.79706 0 1
F2_7 1.46241 -0.61563 0 1
F2_9 1.47568 -0.16101 0 1
F2_11 1.87563 -0.76325 0 1
F2_13 1.59564 -1.27599 0 1
F2_15 2.34207 -0.42311 0 1
F2_17 1.57916 0.63518 0 1
F2_19 2.23561 2.50587 0 1

$means
F1
0

$cov
  F1
F1  1

# obtain person estimates via fscores & display first 6 cases
> head((peopleforms1_2=fscores(concurrent,method="EAP",full.scores=T,full.scores.SE=T)),6)
      F1      SE_F1
[1,] -0.2878288 0.2831983
[2,]  0.4101947 0.2839257
[3,] -1.9938284 0.5392805
[4,]  0.4264012 0.2847325
[5,] -0.9811770 0.3447239
[6,] -0.7943341 0.3219926

> tail(peopleforms1_2,4)
      F1      SE_F1
[1997,]  0.5716681 0.3087964
[1998,]  0.4642074 0.3020029
[1999,]  0.7636787 0.3240319
[2000,] -0.3535776 0.2940903

> mean(peopleforms1_2[,1])
[1] -0.0003883485

> sd(peopleforms1_2[,1])
[1] 0.9403164

> # person estimates written to output file:
> write.csv(peopleforms1_2, file = "peopleforms1_2.csv")

```

(continued)

TABLE 11.8. (continued)

	a	b	g	u
F1_1	1.68062	0.86362	0	1
F1_3	2.59350	0.24188	0	1
F1_5	1.91138	0.76321	0	1
F1_7	1.74657	-0.45098	0	1
F1_9	1.51776	-0.03662	0	1
F1_11	1.97823	-0.70848	0	1
F1_13	1.71845	-1.29787	0	1
F1_15	2.16939	-0.33143	0	1
F1_17	1.41970	0.71320	0	1
F1_19	2.17940	2.62724	0	1
C_2	1.71877	-0.24805	0	1
C_4	1.90026	0.13389	0	1
C_6	2.05406	-0.23107	0	1
C_8	1.94630	0.35342	0	1
C_10	1.97787	-0.58681	0	1
C_12	2.10332	0.36986	0	1
C_14	1.51788	0.90744	0	1
C_16	1.45132	-1.13402	0	1
C_18	1.30843	1.72436	0	1
C_20	1.73872	0.44232	0	1
F2_1	1.23190	1.11345	0	1
F2_3	1.86142	0.25301	0	1
F2_5	1.71839	0.79706	0	1
F2_7	1.46241	-0.61563	0	1
F2_9	1.47568	-0.16101	0	1
F2_11	1.87563	-0.76325	0	1
F2_13	1.59564	-1.27599	0	1
F2_15	2.34207	-0.42311	0	1
F2_17	1.57916	0.63518	0	1
F2_19	2.23561	2.50587	0	1

is referred to as linking. Equating is used when different groups of people are on different metrics and we wish to place all the individuals on a common metric. Therefore, the focus of linking is to adjust item parameter estimates, and the focus of equating is to adjust person location estimates. Placing the items or the individuals on a single metric allows us to make comparisons among the items or the individuals. The equating of multiple forms assumes that all the forms are measuring the same construct and that these forms have been created to the same content and statistical specifications.

The linking and equating processes consist of a data collection phase followed by a transformation phase. There are multiple data collection approaches, such as the single group with counterbalancing method and the common-item nonequivalent groups method. Once the data are collected, then we apply a transformation procedure to align the different metrics. These transformation procedures may be classified as traditional approaches (e.g., linear equating, equiprocentile equating) or modern approaches (e.g., total characteristic function equating).

In the context of IRT, the transformation procedure requires determining the metric transformation coefficients ζ and κ to transform the item parameters and person parameters (or their estimates). Two common approaches for determining the metric

transformation coefficients are linear equating and total characteristic function equating. The former uses the standard deviations and means of the item location parameters (or their estimates), whereas the latter uses all the item parameters (or their estimates) to determine the metric transformation coefficients. In general, if we are using a model with a constant discrimination parameter, then linear equating may be a useful approach. However, whenever item discrimination is allowed to vary across items, then the total characteristic function approach is the preferred technique.

In the next chapter we revisit a model–data fit issue, differential item functioning (DIF). DIF occurs when different groups of individuals perform differently on an item even after we control for or take into account differences in the latent variable, θ . In short, individuals in one group can have a more difficult time endorsing a particular response than those in the other group for reasons that have nothing to do with their positions on the latent variable. There are multiple approaches for identifying the presence of DIF. Some of these approaches require that the item parameter estimates from the different groups first be linked before applying the DIF method. We discuss both non-IRT- and IRT-based methods.

Notes

1. The process discussed is sometimes referred to as *horizontal equating*. In horizontal equating, we are interested in adjusting person location scores for slight differences in, for example, difficulty across forms. If the procedure is successful in disentangling form differences from person differences, then the equated forms are *interchangeable*. A related procedure is *vertical scaling*. (Although this procedure is sometimes called *vertical equating*, the preferred term is vertical scaling.) In vertical scaling, we are interested in creating a common metric for comparing individuals across different educational levels (e.g., grades). Because examinations are developed to be appropriate for a specific education level, they are most likely inappropriate for other education levels. As such, the instruments are *not interchangeable* even after the procedure is completed. For this reason, the term *equating* is reserved for those techniques that result in an *interchangeability* of forms. Vertical scaling is considered to be a method for *scaling to achieve comparability* (AERA, APA, NCME, 1985). In situations involving children and a wide range of education levels, developmental changes in the children may adversely affect the creation of a meaningful common scale. Petersen et al. (1989) discuss issues that are relevant to vertical scaling in the context of grade-equivalent scales; also see Baker (1984). A related issue is scale shrinkage (Yen, 1985). As originally defined, *scale shrinkage* refers to a pattern of increasing mean discrimination and decreasing item/person location variability as educational level increases after the tests have been vertically scaled. Camilli (1988, 1999) differentiates between within-education level shrinkage and between-education levels shrinkage. In both cases, one sees the pattern of decreasing variances across equated/scaled tests. Changes in dimensionality, reliability, and scaling procedures have been invoked to explain the occurrence of scale shrinkage.

2. The preceding assumes that all administered items make up the *operational* set of items. In the ability/achievement context, additional item sets may be administered to the examinees. The examinees' performance on these additional sets is not used in determining their scores on the instrument. These *nonoperational* item sets are included to gather information on the items' characteristics as part of pretesting the items, either for their calibration or as part of a preequating design. For more information on preequating, see Kolen and Brennan (1995, 2004) or Lord (1980).
3. A number of additional situations may lead to metrics being unaligned. For example, because different programs may use different approaches for resolving the indeterminacy issues (e.g., person vs. item centering), the corresponding estimates are on different metrics even when the same calibration sample is used. A second example involves the creation of an item bank from which we would develop multiple forms. For security reasons we would continuously be developing new items for addition to the bank. Because these new items would most likely be administered to a different group of individuals than those used in creating the item bank, these new items would not be on the item bank's metric. Rather, the new items would be on the metric defined by the individuals used in their calibration, and this metric would have to be linked to that of the item bank. However, any instruments developed from the item bank will yield estimated person locations that are on the same continuum as the item bank.
4. Haebara's (1980) method differs from the Stocking and Lord approach by using a slightly different criterion function that is focused on the IRFs from the two metrics. In Haebara's method, the transformation coefficients are those values that minimize the differences between corresponding IRFs across the samples. That is, criterion function is item-level focused rather than on Stocking and Lord's instrument-level focus. Specifically, let θ_{12}^* θ_{21}^* represent the transformation of a person location from Form 1 (initial metric) to Form 2 (target metric) and θ_{21}^* be the reverse transformation of a person location on Form 2 to Form 1 (i.e., $\theta_{12}^* = \zeta\theta_1 = \kappa$ and $\theta_{21}^* = (\theta_2 - \kappa)/\zeta$). Because a perfect equating from Form 1 to Form 2 implies $p_{j,1}(\theta_1) = p_{j,2}(\theta_{12}^*)$, then the discrepancy between $p_{j,1}(\theta_1)$ and $p_{j,2}(\theta_{12}^*)$ reflects equating error in transforming person j 's location on Form 1 to Form 2 (i.e., $e_{ij,1} = p_{j,1}(\theta_1) - p_{j,2}(\theta_{12}^*)$). Haebara subdivides the θ_1 continuum into n_1 intervals, with the interval midpoints used for calculating the probabilities. The n_1 relative frequencies (rf_1) serve as weights for the errors. Summing the weighted squared errors across items and the n_1 intervals gives us the first component of his loss function.

The loss function's second component takes into account the equating error effect of Form 2's $\theta_{i,2}$ s. These errors are defined as $e_{ij,2} = p_{j,2}(\theta_2) - p_{j,1}(\theta_{21}^*)$. Similar to what was done with Form 1's continuum, we subdivide the θ_2 continuum into n_2 intervals with the interval midpoints used for calculating the probabilities. As above, the relative frequencies (rf_2) of the n_2 intervals are used as weights.

Putting our two components together, we arrive at Haebara's loss function

$$F = \sum_{j=1}^L \left[\sum_{i=1}^{n_1} e_{ij,1}^2 rf_1(\theta_{i,1}) + \sum_{i=1}^{n_2} e_{ij,2}^2 rf_2(\theta_{i,2}) \right]. \quad (11.13)$$

As with the Stocking and Lord approach, the values of κ and ζ that minimize F are the transformation coefficients. Other criteria have been suggested. For instance, Divgi (1985) suggested a different criterion that could be solved using a minimum chi-square approach.

5. An alternative approach to obtaining the item parameter estimates from the output object is to use the `extract.item` function. Recall from Chapter 4 that using the `extract.item(. . . ,i)@par` function yields estimates in the slope–intercept format for the four-parameter model. In our example, form 2's value are

```
> for(i in 1:nitems){
+ form2est[i,]=(extract.item(form2,i)@par)    }
> colnames(form2est)=c('alpha','gamma','chi','e') # meaningful
                                                 variable names
> form2est
      alpha      gamma      chi   e
1  1.158759 -1.2524647 -999 999
2  1.565799  0.6601013 -999 999
3  1.751338 -0.2879585 -999 999
4  1.689571 -0.0953457 -999 999
5  1.614081 -1.2025826 -999 999
6  1.845988  0.6841702 -999 999
7  1.380624  1.0459674 -999 999
8  1.784986 -0.4347550 -999 999
9  1.391003  0.3834233 -999 999
10 1.740562  1.4468454 -999 999
11 1.779921  1.6251275 -999 999
12 1.890414 -0.5800473 -999 999
13 1.516890  2.2008279 -999 999
14 1.312463 -1.1733487 -999 999
15 2.214095  1.2301084 -999 999
16 1.238355  1.7543225 -999 999
17 1.484968 -0.8493459 -999 999
18 1.073460 -2.0180606 -999 999
19 2.118051 -5.4077561 -999 999
20 1.543531 -0.5118222 -999 999
```

SNSEquate assumes a logistic deviate format. Consequently, the intercepts (γ) need to be transformed to locations by $\hat{\delta}_j = -\hat{\gamma}_j/\hat{\alpha}_j$ (see Equations 2.5 and 5.1). For instance,

```
> form2est$gamma = -1*(form2est$gamma/form2est$alpha)
```

Moreover, the `e` variable needs to be removed using the `rm` function.

12

Differential Item Functioning

As previously mentioned, when we have model–data fit, our parameter estimates are invariant across samples. Accordingly, the presence or absence of invariance is indicative of model–data fit. To this end, we divide our calibration sample into two random subsamples, we calibrate each subsample, and we compare each subsample’s estimates to one another to examine item parameter estimate invariance (see Chapters 4 and 5). We take, for example, high linearity across subsamples as evidence of invariance. In this chapter, we revisit this idea, albeit using manifest groups of respondents rather than random subsamples. Essentially, the question at hand is the following: “Are items “behaving” differently across our manifest groups?” In other words, do we have invariance across our manifest groups? If the answer is “no,” then the item is “behaving” or functioning differently across our manifest groups.

By way of an example, assume that we have developed the following item to assess general vocabulary knowledge.

What does *alto* mean?

- a. again
- b. also
- c. countertenor
- d. high
- e. in addition

If our respondents to this item were to be divided into Hispanic and non-Hispanic manifest subgroups, we might find that Latinas/os might select option *d*, whereas non-Latinas/os might select option *c*, even after we controlled for differences in vocabulary proficiency. (In fact, some Latinx of Mexican origin may think that the best answer, *stop*, is not even provided.) Therefore, performance on this item is a function not only of the respondent’s English vocabulary proficiency, but also of a tangential variable—the respondent’s ethnic group (specifically, the person’s Spanish proficiency).

Various approaches have been developed to try to identify items that function differently across groups. In the following we discuss some of these approaches. These differential item functioning techniques should be considered a standard part of our toolbox of model–data fit methods. We begin with a discussion of item bias and how it relates to differential item functioning, after which we present three techniques for performing differential item-functioning analyses. Parallel to the structure of previous chapters, we then provide a demonstration of a differential functioning analysis.

Differential Item Functioning and Item Bias

Although the term *bias* has a statistical interpretation (i.e., the systematic under- or overestimation of a parameter), in the layperson's mind bias is typically associated with the issue of unfairness—in other words, an instrument that has an adverse impact on different ethnic or racial groups. As such, the terms *item bias* and *test bias* have certain culturally negative connotations. Despite efforts to disentangle these connotations from the term (e.g., Jensen, 1980), such perceptions have continued. Psychometrically, the definition of bias has evolved from instrument-focus to item-focus. Camilli and Shepard (1994), Holland and Wainer (1993), and Zumbo (2007) present the history of bias in testing as well as approaches for the analysis of bias scores.

The current practice for determining whether an instrument is biased is to examine the instrument at the item level to see whether one or more items may be considered biased. To identify items as biased involves using *differential item functioning* (DIF) methods to detect items that are functioning differently across manifest groups of individuals (e.g., Hispanics and non-Hispanics). An item identified as exhibiting DIF is reviewed by a panel of experts to determine whether the source of an item's differential performance is relevant or irrelevant to the construct being measured; this review is also known as "logical evidence of bias." It is the panel's conclusion that determines whether an item exhibiting DIF is also considered biased. In practice, items are subjected to sensitivity reviews to remove material that may be considered offensive or demeaning to particular groups, regardless of DIF analyses (Camilli & Shepard, 1994). For example, the term *colored people* is considered offensive in the United States, but not necessarily in other countries (i.e., different cultures). As such, the term would normally not be used in an item.

In the following discussion, we focus on methods for identifying DIF. Although our presentation revolves around proficiency assessment, this should not be interpreted to mean that DIF is a concern only in proficiency assessment. For instance, attitude and personality inventories may contain items that require that the respondents have certain knowledge or a particular background in order to understand the items as intended. However, because of ethnic/racial, gender, and/or cultural differences in the respondents, this may not be true.

DIF is defined as an item that displays different statistical properties for different manifest groups after the groups have been matched on a proficiency measure (Angoff, 1993). For instance, assuming binary data and an IRT framework, we find that DIF is

reflected as a difference between the conditional probabilities of a response of 1 for two manifest groups (e.g., females and males). In the DIF nomenclature, one of the manifest groups is known as the *Focal group*, whereas the other is called the *Reference group*. The Focal group (e.g., females) is the one being investigated to see if it is disadvantaged by the item. The Reference group is the comparison group (e.g., males). In some item bias literature, the Focal group is called the “minority” (membership) group and the Reference group is the “majority” (membership) group.

Graphically, DIF can be represented as the difference between two IRFs. One IRF is based on the item’s parameter estimate(s) from the Focal group, whereas the other IRF is based on the item’s parameter estimate(s) from the Reference group. If an item is not exhibiting DIF, then the groups’ IRFs would be superimposed on one another (i.e., within sampling error) after we link the two groups’ metrics. However, if the item is exhibiting DIF, then the two IRFs are not superimposed after we link the two groups’ metrics. Figure 12.1 presents an example of an item exhibiting DIF.

The item shown in Figure 12.1 favors members of the Reference group (solid line) over those from the Focal group (dashed line). In other words, we see that throughout the θ continuum, the probability of a response of 1 is higher for Reference group members than for Focal group members. This form of DIF is known as *uniform DIF*. If this is a proficiency item, then the item is easier for members of the Reference group than for Focal group members.

The second type of DIF is *nonuniform DIF*. With this type, members of the Reference group perform better than Focal group members for part of the θ continuum, but this

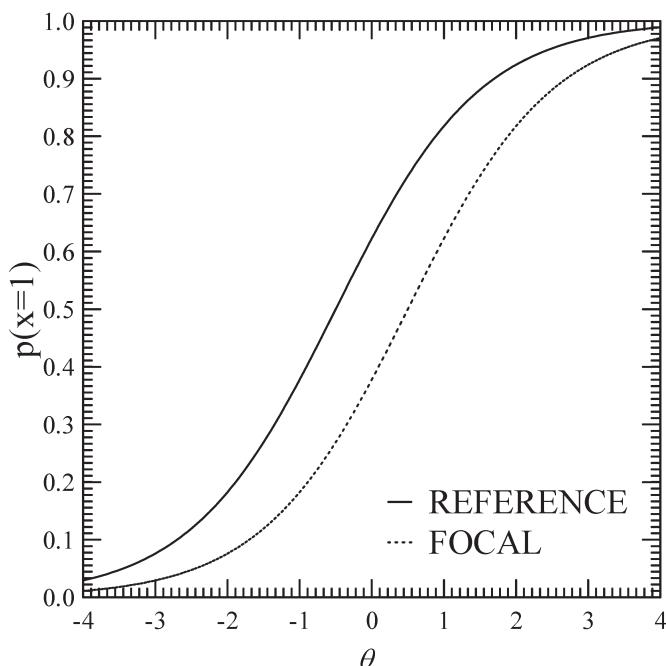


FIGURE 12.1. Example of uniform differential item functioning for one item.

relationship is reversed for a different part of the continuum. Graphically, an item that exhibits nonuniform DIF has IRFs that cross. Figure 12.2 contains an example of an item exhibiting nonuniform DIF. Above $\theta = 0$ Reference group members have a higher probability of responding with a 1 than do members of the Focal group. However, below $\theta = 0$ Focal group members have a higher probability of responding with a 1 than do members of the Reference group. With nonuniform DIF it is possible that DIF for one group (positive DIF) may be wholly or partially compensated for by DIF against that group (negative DIF) at another point along the latent variable continuum. To summarize, in nonuniform DIF an item's IRFs differ in their slopes and potentially their locations, whereas with uniform DIF an item's IRFs differ only in their locations.

One explanation for why an item exhibits DIF is based on multidimensionality. That is, DIF can be conceptualized as a form of multidimensionality that occurs when an item measures multiple dimensions and when the manifest groups differ in their relative locations to one another on the nonprimary latent variable(s). If the two groups do not differ in their relative locations on the nonprimary dimension(s), then neither group benefits from the nonprimary dimension, nor does DIF occur even though the data are multidimensional (cf. Ackerman, 1992).

There are three additional issues we need to discuss. As noted above, once an item is identified as exhibiting DIF, its text is subjected to panel review to determine if the wording of the item may explain the differential performance. If an acceptable explanation is not forthcoming, then the item is not considered biased, although it may be elimi-

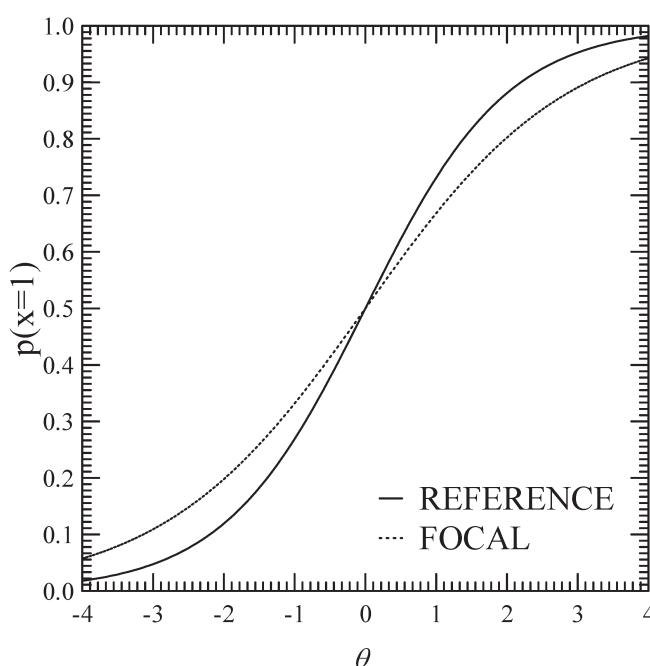


FIGURE 12.2. Example of nonuniform differential item functioning for one item.

nated from the instrument for exhibiting DIF. In short, flagging an item as exhibiting DIF is a necessary but not sufficient condition for the item being considered to be biased.

The second issue is the degree of DIF. As we know, in hypothesis testing a statistically significant result does not necessarily mean the result is meaningful. This concept may be applied to DIF analysis. For instance, the Educational Testing Service (ETS) classifies items exhibiting DIF into three categories, depending on the significance of the test statistic and the magnitude (i.e., degree) of the DIF (Dorans & Holland, 1993; Zieky, 1993).

The third issue is the impact of an item exhibiting DIF on the DIF analysis itself. In other words, because DIF is defined as a difference between conditional probabilities, a DIF analysis conditions on the respondents' location estimate. Consequently, what is the impact of one or more DIF items on the conditioning variable used in the DIF analysis? Camilli and Shepard (1994) strongly recommend that the items exhibiting the greatest DIF and that have been determined to be biased be removed from the calculation of the conditioning variable and that the DIF analysis be redone. By using this iterative process, the conditioning variable is "purified." The procedure's premise is that using this purified conditioning variable will identify the same items as exhibiting significant DIF, while also helping to reveal whether other DIF items were masked by the contaminated conditioning variable. Holland and Thayer (1988) conjectured "that it is correct to *include* the studied item in the matching criterion when it is being analyzed for DIF, but if it has substantial DIF then that item should be *excluded* from the criterion used to match examinees for any *other* studied item" (p. 143). Their reasoning is that including the item controls the size of the test, whereas removing the item prevents large DIF items from reducing the power of the test.

From an IRT perspective, the existence of DIF means the DIF item's parameter estimates are not invariant across the manifest groups (i.e., item–data misfit).¹ As mentioned above, it is the lack of invariance that is interpreted as evidence of DIF. Various IRT-based indices have been created for DIF detection. Among the DIF approaches based on IRT are the likelihood ratio ($TSW-\Delta G^2$) method outlined by Thissen, Steinberg, and Wainer (1988, 1993); Lord's Chi-Square (Lord, 1980); and the Exact Signed Area and H Statistic approaches (Raju, 1988, 1990); see Kim and Cohen (1995) and Hidalgo and López-Pina (2004) (to name just two studies) for a performance comparison.² Non-IRT-based approaches include, but are not limited to, the nonparametric Mantel–Haenszel Chi-Square (MH) statistic (Holland & Thayer, 1988), log linear modeling (Mellenbergh, 1982), multilevel modeling (e.g., French & Finch, 2010), and logistic regression (Swaminathan & Rogers, 1990).

In the next section, we discuss three approaches for DIF detection: the MH statistic, the $TSW-\Delta G^2$, and the logistic regression approach. Each approach is implemented by examining each item on an instrument for DIF. Although we do not discuss Lord's Chi-Square or Raju's Exact Signed Area and H Statistics, it should be noted that these DIF indices require that the groups' metrics be linked prior to the technique's application; this is done by using one of the approaches discussed in Chapter 11. In the following discussion, we assume the Reference group is coded 0, the Focal group is coded 1, and that we are working with dichotomous response data.

Mantel-Haenszel Chi-Square

The Mantel-Haenszel statistic (MH; Mantel & Haenszel, 1959) is used for determining whether two variables are independent of one another while conditioning on a third variable (i.e., the analysis of a three-way contingency table). As applied to DIF detection, the MH statistic consists of the sum of a series of 2×2 contingency tables where each table is based only on persons who received the same summed (or summated) score. Typically, the conditioning variable is the summed score or some modification of the summed score. Each table is the cross-classification of an item's binary responses with group membership. Because for a zero summed score or a perfect summed score the 2×2 table collapses to have a single column that represents responses of 0 or 1, respectively, there are always one fewer tables than the maximum possible summed score on the instrument. For example, assuming a four-item instrument with summed scores of 0 to 4, we have three 2×2 tables: one 2×2 table is for a summed score of 1, another is for a summed score of 2, and a third table is for a summed score of 3.

To calculate the MH statistic, we need to create each of the 2×2 tables conditioned on the summed scores. Let t represent the t th summed score where $t = 1, \dots, (L-1)$, and L is the instrument's length. Then, for the item of interest and the t th summed score, the following 2×2 table of manifest groups by item response can be formed:

Manifest groups	Item response		<i>Total</i>
	0	1	
Reference	B_t	A_t	n_{Rt}
Focal	D_t	C_t	n_{Ft}
<i>Total</i>	n_{0t}	n_{1t}	n_t

where A_t , B_t , C_t , and D_t are the frequencies for the corresponding cells, n_{Rt} and n_{Ft} are, respectively, the number of Reference and Focal group respondents that obtained the t th summed score, n_{1t} is the number of responses of 1, n_{0t} is the number of responses of 0, and n_t is the number of respondents in t th table. Because we create a 2×2 table for each of the remaining summed scores, each table is conditioned on a summed score.

The MH statistic allows us to determine if the responses to an item are independent of group membership after conditioning on the summed scores. Therefore, the null hypothesis is one of conditional independence. Following Holland and Thayer (1988), we can calculate MH by

$$\text{MH } \chi^2 = \frac{\left(\left| \sum_{t=1}^{L-1} (A_t - E(A_t)) \right| - 0.5 \right)^2}{\sum_{t=1}^{L-1} \text{var}(A_t)}, \quad (12.1)$$

where

$$\text{var}(A_t) = \frac{n_{Rt} n_{Ft} n_{It} n_{Ot}}{n_t^2 (n_t - 1)}$$

and the conditional expectation of A_t is that the responses and group membership are independent for the t th summed score

$$\varepsilon(A_t) = \frac{n_{Rt} n_{It}}{n_t}$$

The value of 0.5 in Equation 12.1 is Yates's correction for continuity.

MH χ^2 is evaluated against the standard χ^2 critical values with degrees of freedom equal to 1. A complementary way of interpreting the null hypothesis is that the odds of Reference group members responding 1 are the same as those of Focal group members. This is referred to as the constant odds ratio hypothesis. Symbolically, the constant odds ratio is represented as α_{MH} and the null hypothesis is $\alpha_{MH} = 1$. A significant MH χ^2 indicates that the odds for the two groups vary across some or all of the 2×2 tables for the item in question. In short, the item is exhibiting DIF.

Examination of Equation 12.1 shows that if the frequency of the Reference group receiving a response of 1 (i.e., A_t) is consistently greater than what we would expect or consistently less than what we would expect for each of the 2×2 tables, then we obtain a positive statistic. In the former case, one has uniform DIF in favor of the Reference group, whereas in the latter case the uniform DIF favors the Focal group. However, if for some of the 2×2 tables A_t is sometimes greater and for other tables it is sometimes less than what we would expect, then the MH χ^2 can approach zero, owing to cancellation occurring across the $(L - 1) 2 \times 2$ tables. In other words, for some of the 2×2 tables (e.g., for some proficiency levels), the Reference group does better than expected, but this is negated by the Focal group doing better than expected for other tables. When this occurs, one has an interaction between group membership, item performance, and the conditioning variable, and this is evidence of nonuniform DIF. In this situation, the MH χ^2 statistic may be nonsignificant, reflecting a violation of the assumption that the direction and degree of association between our manifest groups and responses are the same in 2×2 tables. The MH χ^2 is typically seen as being able to detect only uniform DIF, although in some cases one may obtain a significant statistic when nonuniform DIF. Although this is possible, we are not suggesting that the MH χ^2 statistic be used for detecting nonuniform DIF, but rather that a significant MH χ^2 statistic does not necessarily mean the presence of only uniform DIF.

To summarize, the MH χ^2 statistic determines whether a relationship exists between performance on an item and group membership, after taking into account performance on the instrument. Stated another way, the statistic determines whether the odds of success for Focal group members significantly differ from the odds of success for comparable Reference group members across the $(L - 1) 2 \times 2$ tables. However, the test does not give us an idea of the strength of this relationship. To obtain an indication of the degree

of association, we can estimate the common odds ratio across the $(L - 1)$ 2×2 tables. This estimate may be calculated by (Mantel & Haenszel, 1959)

$$\hat{\alpha}_{MH} = \frac{\sum_{t=1}^{L-1} \left(\frac{A_t D_t}{n_t} \right)}{\sum_{t=1}^{L-1} \left(\frac{B_t C_t}{n_t} \right)} \quad (12.2)$$

When $\hat{\alpha}_{MH}$ equals 1, then on average, the odds for Focal group members responding 1 are the same as the odds for comparable Reference group members on the studied item (i.e., no DIF on the item). When $\hat{\alpha}_{MH}$ is greater than 1, then, on average, the Reference group members performed better than comparable Focal group members on the item of interest (Holland & Thayer, 1988). For instance, if $\hat{\alpha}_{MH} = 2$, then Reference group members, on average, have twice the odds of success as comparable Focal group members. Conversely, when $\hat{\alpha}_{MH}$ is less than 1 then, on average, the Reference group members performed worse than comparable Focal group members. It should be noted that the individual odds ratios that are summed to obtain this common odds ratio should not vary drastically from one another (Agresti, 1996)—that is, we have homogeneous odd ratios.

Although $\hat{\alpha}_{MH}$ indicates how much better or worse on average the Focal group members performed relative to comparable Reference group members, its scale is asymmetric with a lower bound of 0, an upper bound of ∞ , and no DIF indicated by a value of 1. Therefore, as is typically done with odds and odds ratios, α_{MH} is transformed to a scale symmetric about 0. Specifically, we can transform α_{MH} to the natural logarithmic scale to obtain a log odds ratio, $\beta_{MH} = \ln(\alpha_{MH})$, where a 0 indicates no DIF (cf. Agresti, 1996; Camilli & Shepard, 1994). A second transformation found in the literature (e.g., Holland & Thayer, 1988) is to further transform the log odds ratio to be on the ETS difficulty delta scale by $D_{MH} = -2.35 \beta_{MH} = -2.35 \ln(\alpha_{MH})$.

The magnitude of the log odds ratio β_{MH} (or the corresponding D_{MH}) indicates the degree of DIF on the item. Conceptually, this may be represented by how far away the Focal group IRF is from the Reference group IRF. A $\beta_{MH} = 0$ (or $D_{MH} = 0$) indicates that the two groups perform the same on the item and the corresponding IRFs are superimposed. A positive β_{MH} (or a negative D_{MH}) indicates that, on average, Reference group members tend to provide responses of 1 more often than comparable Focal group members. Therefore, the Reference group IRF is to the left of the Focal group IRF by an amount “equal” to the magnitude of the DIF. In a proficiency testing context, this means that the item is easier for Reference group members than for Focal group members of similar proficiency. In contrast, a negative β_{MH} (or a positive D_{MH}) indicates that, on average, Focal group members tend to provide responses of 1 more often than comparable Reference group members. Consequently, the Reference group IRF is to the right of the Focal group IRF by an amount “equal” to the magnitude of the DIF. That is, the item is easier for Focal group members than for Reference group members.

In addition to the simple point estimate of β_{MH} , it is possible to calculate a con-

fidence interval for $\hat{\beta}_{MH}$ to estimate the range within which the true logit would be expected to fall. Accordingly, the variance error for $\hat{\beta}_{MH}$ is (see Holland & Thayer, 1988)

$$s_e^2(\hat{\beta}_{MH}) = \frac{1}{2 \left[\sum_t \frac{A_t D_t}{n_t} \right]^2} \sum_t \left[n_t^{-2} (A_t D_t + \hat{\alpha}_{MH} B_t C_t) (A_t + D_t + \hat{\alpha}_{MH} (B_t + C_t)) \right]. \quad (12.3)$$

Using the unit normal distribution, we find that our $100(1 - \alpha)\%$ CI = $\hat{\beta}_{MH} \pm z_{(1-\alpha/2)} s_e(\hat{\beta}_{MH})$.

The ETS three-category DIF classification system is based on a combination of the magnitude of D_{MH} and its statistical significance at a 5% significance level. Specifically, the three categories are A (negligible DIF), B (intermediate DIF), and C (large DIF). “A” items have a nonsignificant D_{MH} with respect to 0 or have a $D_{MH} < 1$, whereas “C” items have $|D_{MH}| > 1.5$ and is statistically larger than 1. All items not classified as “A” or “C” are “B” items (Dorans & Holland, 1993). The statistical significance of D_{MH} is evaluated using a modified version of Equation 12.3 in which n_t^{-2} becomes $2n_t^{-2}$ and $1/2 \left[\sum_t \frac{A_t D_t}{n_t} \right]^2$ becomes $2.35 \left[\sum_t \frac{A_t D_t}{n_t} \right]$ (Dorans & Holland, 1993 DIF).

The TSW Likelihood Ratio Test

The Thissen, Steinberg, and Wainer (1988) DIF detection strategy is based on a comparison of the fit of two IRT models using the likelihood ratio test statistic introduced in Chapter 6. This comparison determines if there is a significant difference in model fit when one constrains an item to have the same location across groups versus when the item is free to have different locations across the manifest groups. Ideally, when we allow the item location to vary across the groups, the two location estimates should be identical. This would indicate that the item is performing the same way in both groups. Thus, the fit would be the same, and we would observe a nonsignificant likelihood ratio test for the item in question. The null hypothesis tested by the likelihood ratio test (TSW- ΔG^2) is that there are no group differences in the item parameter estimates. One may also choose to include the discrimination parameter to simultaneously test whether this varies across groups.

Implementing the TSW- ΔG^2 approach is a three-step procedure. As an example, assume we are investigating whether item 1 on an instrument is exhibiting DIF. In step 1 we would fit an IRT model, such as the 1PL model, to both manifest groups with the proviso that item parameter estimates for all items, except for item 1, be constrained to be equal across groups. As a result, it is possible for item 1 to have different location estimates across the two groups. For step 2 we fit the same IRT model, but this time all item parameter estimates, including those of item 1, are constrained to be equal across both groups; this is the “no DIF” or null situation. Step 3 is the calculation of TSW- ΔG^2

$$\text{TSW-}\Delta G^2 = G_2^2 - G_1^2, \quad (12.4)$$

where G_1^2 and G_2^2 are the likelihood ratios from steps 1 and 2, respectively. TSW- ΔG^2 is distributed as a χ^2 (when the sample size is large) with degrees of freedom equal to the number of item parameters allowed to differ across the groups and when the nesting model holds for the data. In our example with the 1PL model, δ would be allowed to differ across groups, and the df for evaluating the significance of TSW- ΔG^2 would be 1; for the 2PL model where both α and δ are simultaneously investigated, the df would be 2, and so on. A significant TSW- ΔG^2 indicates the presence of DIF for the item under consideration. Conversely, a nonsignificant TSW- ΔG^2 indicates the item is not exhibiting DIF. This three-step process is repeated for all the items on the instrument.

The TSW- ΔG^2 DIF approach can be performed using IRTPRO, and flexMIRT, difR, IRTLRDIF (Thissen, 2001), and mirt to name a few software options. The gist of the procedure requires one to specify the equality constraints. The $-2\ln L$ (i.e., negative twice the log likelihood) value at the end of the output would be used as either G_1^2 or G_2^2 , depending on which model was being fitted. To perform this analysis, one would subdivide the calibration sample into the two manifest groups. The data file's structure would contain all the responses for the first group in columns 1 through L. Following these responses would be all the responses for the second group, but their responses would begin in the $(L + 1)$ th column and continue to the $2L$ th column. For instance, assuming a five-item instrument, group 1's responses to item 1 would be in column 1 and group 2's responses to item 1 would be in column 6, group 1's responses to item 2 would be in the second column and group 2's responses to item 2 would be in the seventh column, and so on. To impose the equality constraint on item 1, we would refer to items 1 and 6 (i.e., the first item on the instrument is labeled item 1 in group 1 and item 6 in group 2). By using the calibrations' $-2\ln L$ values, it is possible to work with multiple-model calibrations of an instrument (e.g., a mixed item format instrument). Unlike some IRT-based approaches, because with TSW- ΔG^2 both groups' responses reside in one data file, the estimates are on the same metric. (Some IRT-based approaches require linking Reference and Focal group metrics prior to performing the DIF analysis.) Thissen et al. (1993) provide an example of implementing their procedure.

Logistic Regression

Logistic regression is a technique for making predictions about a binary variable from one or more variables. These predictor variables may be quantitative and/or qualitative. In the current context, the binary variable is the response to an item, and the predictors might be gender and/or some measure of the construct. As such, we logically regress the responses to item j on the construct measure and/or on gender.

In Chapter 2, Equation 2.1, we presented a general form of the logistic regression model. By way of an example, the z in Equation 2.1 in a one predictor situation can be written as $\beta_0 + \beta_1 X$, where β_0 is the intercept (or constant), β_1 is the regression coefficient, and X is a predictor (e.g., gender). The model's $p(x)$ may be interpreted as the conditional mean of the criterion, given x , when the logistic distribution is used (Hosmer & Lemeshow, 2000). Furthermore, we may extend z to include multiple predictors

including the interaction of predictors. Therefore, a logistic regression analysis allows us to assess the effect of one or more predictors on the observed responses for an item. Both Hosmer and Lemeshow (2000) and Agresti (1996) contain readable introductions to logistic regression.

Conceptually, the application of logistic regression to DIF analysis requires performing a logistic regression analysis for an item, using members of the Reference group, and a second analysis for the same item with members of the Focal group. The first analysis provides estimates for the constant (β_{0R}) and regression coefficient (β_{1R}) for the Reference group. Similarly, the second analysis estimates the constant (β_{0F}) and regression coefficient (β_{1F}) for members of the Focal group. If the intercept terms are equal (i.e., $\beta_{0R} = \beta_{0F}$) and the regression coefficients are equal (i.e., $\beta_{1R} = \beta_{1F}$), then the predicted probability curves are identical and there is no evidence of DIF. However, if the constants and/or the regression coefficients are unequal, then there is some indication of DIF. If the regression coefficients are equal (i.e., $\beta_{1R} = \beta_{1F}$), but the constants are unequal (i.e., $\beta_{0R} \neq \beta_{0F}$), then the predicted probability curves are separate and parallel to one another. This would represent uniform DIF. Conversely, if the coefficients are unequal (i.e., $\beta_{1R} \neq \beta_{1F}$), then the predicted probability curves cross and there is evidence of nonuniform DIF.

Although conceptually we may view the logistic regression approach for DIF analysis as two separate analyses, in practice, these two analyses are combined into a series of nesting and nested models. As a result, before we present the mechanics of the logistic regression approach we need to supplement the terminology we have used so far.

Recall that the TSW likelihood ratio test compares a model that assumes a common *IRT-based* IRF for both the Reference and Focal groups (i.e., the “no DIF” model), with a second model that allows the IRFs to differ across these groups (i.e., “DIF exists” model). By imposing the equality constraints on the item parameter estimates in the “DIF exists” model, one obtains the “no DIF” model. As a result, the “no DIF” model may be seen as nested within the “DIF exists” model. Moreover, the “no DIF” model may be seen as a *reduced* version of the “DIF” model because it has fewer parameters than does the “DIF” model. Because the “DIF” model subsumes the “no DIF” model, the “DIF” model may be viewed as the *full* model. (Sometimes the reduced model is called the *compact* model and the full model the *augmented* model [see Thissen et al. (1993)]). In Chapter 6 we introduced the likelihood ratio test statistic to determine whether the full model differed significantly from the reduced model

$$\Delta G^2 = -2 \ln \left[\frac{L_R}{L_F} \right] = (-2 \ln L_R) - (-2 \ln L_F), \quad (12.5)$$

where L_R is the maximum of the likelihood for the reduced model and L_F is the maximum of the likelihood for the full model; we present the first form of ΔG^2 to show the “ratio” in the likelihood ratio test statistic.³ The *df* for evaluating the significance of ΔG^2 are the difference in the number of parameters in the full and the reduced models.

In the context of a DIF analysis, the full model is defined as (cf. Swaminathan & Rogers, 1990)

$$\hat{z} = \beta_0 + \beta_1 \Lambda + \beta_2 \Gamma + \beta_3 (\Lambda * \Gamma), \quad (12.6)$$

where Λ is a measure of an individual's position on the latent variable (e.g., Λ may be $\hat{\theta}$ or X), Γ is a categorical predictor variable indicating group membership for an individual ($\Gamma = 1$ for members of the Focal group and $\Gamma = 0$ for members of the Reference group), and $(\Lambda * \Gamma)$ is the interaction of a person's location on the latent variable and their group membership.⁴ Although Γ may be a manifest or latent categorical or continuous variable, in the following we treat it as a categorical manifest variable to be consistent with the traditional application of logistic regression to DIF analysis. (See Appendix G, "Should DIF Analyses Be Based on Latent Classes?" for an alternative conceptualization.) The term β_1 indicates the relationship between the performance on the item and the person's location on the latent variable, β_2 corresponds to the mean group difference in performance on the item (i.e., $\beta_2 = \beta_{0F} - \beta_{0R}$), and β_3 reflects the group by person location interaction (i.e., $\beta_3 = \beta_{1F} - \beta_{1R}$). Our estimates of β_0 , β_1 , β_2 , and β_3 , are b_0 , b_1 , b_2 , and b_3 , respectively.

With respect to the full model, there are two reduced models to consider. The first reduced model omits the interaction term from Equation 12.6. In this case, the reduced model (1) is the uniform DIF model

$$\hat{z} = \beta_0 + \beta_1 \Lambda + \beta_2 \Gamma. \quad (12.7)$$

A comparison of reduced model (1) with the full model (Equation 12.6) tests to see whether the interaction term is necessary to account for response variability on the item of interest. In other words, ΔG^2 is used to determine whether we should retain or reject the null hypothesis that $\beta_3 = 0$. If we obtain a nonsignificant result (i.e., we retain the null hypothesis), then the model does not need the interaction term given the other terms in the other model. Stated another way, we do not have evidence of nonuniform DIF for the item under consideration. Conversely, if we obtain a significant result, then we have evidence supporting the existence of nonuniform DIF and should retain the interaction term in the model; this conclusion holds regardless of β_2 's value.

We can create a second reduced model by setting $\beta_2 = 0$ in Equation 12.7; Equation 12.7 is a full model. Consequently, our reduced model (2) is the no DIF model

$$\hat{z} = \beta_0 + \beta_1 \Lambda. \quad (12.8)$$

A comparison of reduced model (2) with Equation 12.7 determines whether group membership is necessary to explain performance on the item given the other terms in the model. That is, ΔG^2 is used to test the null hypothesis $\beta_2 = 0$. If we obtain a non-significant result, then the group membership variable may be dropped from the model because the (group) intercept terms are equal. If this is the case, then there would not be evidence of uniform DIF. Conversely, if we obtain a significant result, then this would support the presence of uniform DIF. Implicit in this statement is that we have already determined that there is no evidence of a need for the interaction term (i.e., a comparison of Equations 12.7 and 12.6).

By appropriately formulating Λ , Swaminathan and Rogers (1990) show how the MH

procedure is based on the logistic regression model when Λ is considered to be discrete. Furthermore, the MH log odds ratio, $\beta_{MH} = \ln(\alpha_{MH})$ is equal to Γ 's regression coefficient in Equation 12.7 (i.e., $\beta_2 = \beta_{MH}$). Stated another way, the magnitude of β_2 indicates the difference between the Reference and Focal groups' average performance (in terms of the log odds of success) on the item and β_2 reflects the degree of DIF on the item of interest. A complementary approach has been proposed by Zumbo (1999). Specifically, he proposed using the difference in the models' R^2 s (ΔR^2) to assess the magnitude of DIF. His approach would be useful in the presence of a significant interaction.

There are at least two variants of Zumbo's (1999) approach because of the different types of R^2 statistics in logistic regression. One of these variants is the Nagelkerke R^2 (Nagelkerke, 1991), and a second is the weighted least squares R^2 . Either of these can be used in $\Delta R^2 = R_F^2 - R_R^2$ to assess the DIF effect size in a comparison of the full model (R_F^2) with a reduced model (R_R^2). Because statistical packages like SAS and SPSS (SPSS, 2019) do not calculate the weighted least squares R^2 , one would have to calculate it oneself; Zumbo (1999) presents an example of how to do this in SPSS. Guidelines for what constitutes a negligible, moderate, or large effect size for the weighted least squares R^2 , may be found in Jodoin and Gierl (2001). Specifically, large and moderate DIF are defined as $\Delta R^2 \geq 0.070$ and $0.035 \leq \Delta R^2 < 0.070$, respectively, and a significant test statistic, whereas negligible DIF is defined solely as $\Delta R^2 < 0.035$. Jodoin and Gierl also label large, moderate, and negligible DIF as C-, B-, and A-level DIF items, respectively. In our example below we use Nagelkerke R^2 .

To summarize, to perform the logistic regression DIF analysis, one implements a series of model comparisons using the likelihood ratio (ΔG^2) statistic. First, we compare the models from Equations 12.6 and 12.7 to test for nonuniform DIF (i.e., we test for an interaction term first). If there is evidence of meaningful nonuniform DIF, then the procedure is finished and the item is subjected to panel review to determine whether the item is biased. However, if there is no evidence of meaningful nonuniform DIF, then we proceed to test for uniform DIF by comparing the models in Equations 12.7 and 12.8. Again, if there is evidence of meaningful uniform DIF, then the item is submitted for review. These steps are analogous to not interpreting the main effects in a two-way ANOVA until after we have determined whether we have a significant interaction. Alternatively, we can directly compare the model in Equation 12.8 with the one in Equation 12.6 and perform a 2 df test of the presence of uniform or nonuniform.

The gist of this model comparison series is to find the simplest model that describes the data. As is the case with other DIF methods, the logistic regression technique is applied item by item. The null hypotheses tested are (1) $\beta_3 = 0$ (i.e., nonuniform DIF does not exist) and (2) $\beta_2 = 0$ (i.e., uniform DIF does not exist). Each statistical test is a one degree of freedom comparison. (For the 2 df test the null hypothesis is $\beta_2 = 0$ and $\beta_3 = 0$.) The null hypothesis of $\beta_1 = 0$ is not of particular interest because any reasonably well-constructed instrument will yield person locations that have a relationship to the odds of a response of 1. In effect, the test of $\beta_1 = 0$ is a test of a no predictor (null) model. Although the above describes the technique's application to a binary response item, Zumbo (1999) discusses the technique's application to polytomous ordinal responses (also see French & Miller, 1996).

Example: DIF Analysis of Vocabulary Test, SAS CMH

We demonstrate a DIF analysis by first using the MH statistic, followed by the logistic regression approach. (In Chapter 13 we use a multilevel approach.) We adopt these approaches because of their simplicity, efficacy, and, in the case of MH, their popularity. Moreover, neither one requires us to link Reference and Focal group metrics as would be the case with some of the IRT-based approaches (e.g., Lord's Chi-Square, Exact Signed Area, H Statistic, RMSD [see Chapter 5]). In approaches that use the $\hat{\theta}$ s as the conditioning variable, the Reference and Focal group $\hat{\theta}$ s need to be equated. As such, by using the logistic regression and MH DIF procedures based on X , we avoid introducing linking/equating error into the DIF analysis. Additionally, neither approach is adversely affected by item parameter estimation error and/or errors in $\hat{\theta}$ s.

Our demonstration utilizes data from a study by Subkoviak, Mack, Ironson, and Craig (1984). In their study, items that were expected to exhibit DIF were explicitly incorporated into an examination. Subkoviak et al. administered a 50-item four-option multiple-choice vocabulary test. The test consisted of 40 items that were drawn from the verbal section of the College Qualification Test and involved standard English vocabulary and 10 "Black slang" (B) items. Each item presented a word whose definition the examinee was to choose from one of the four options. Subkoviak et al. provided the following example of a B item; its correct answer is identified by an asterisk:

Greasing A. cleaning *B. eating C. arguing D. talking

Subkoviak et al. hypothesized that the B items would exhibit DIF. Specifically, they believed that African Americans would find the B items to be easier than would Caucasian examinees. The 10 B items were randomly inserted in each block of 5 items on the test. The participants were 1008 African American and 1022 Caucasian students from two universities. The former originated from primarily urban areas and attended a predominantly African American university, whereas the latter were from the rural Midwest and attended a predominantly Caucasian university. For our analysis we modify the response data to consist of 10 items, 2 of which are B items. Unless otherwise indicated, we code the Reference group 0 and the Focal group is coded 1.

For our example, the Reference group represents the Caucasian students (i.e., the "majority" group), and the African Americans students are the Focal group (i.e., the "minority" group). Therefore, given Subkoviak et al.'s hypothesis, our demonstration item should favor members of the Focal group. We use the examinees' summed (i.e., observed) scores as a proxy of their locations on the vocabulary latent variable. In practice, a DIF analysis would be performed using multiple combinations of Reference and Focal groups. That is, first the Reference and Focal groups would be Caucasian and African Americans, then the groups would be male and female, and so on. We begin by using SAS to detect DIF in one B item. Subsequently, we use R to perform the MH analysis.

The SAS program for performing the MH analysis is presented in Table 12.1. The data consist of a case identification field (*person*), the respondent's race (*race*: 0 = Caucasian, 1 = African American) and the binary responses to items 1 through 10

TABLE 12.1. SAS Program for Performing MH Procedure on One Item

```

title "MH DIF example analysis. white reference grp (0), black focal group (1)";
data d1;
  infile "C:\vocab.dat";
  input person race i1-i10;
  X=sum (of i1-i10);
proc freq;
  tables x*race*i3/CMH nopol;

```

(i_1, i_2, \dots, i_{10}). We use the `sum` statement to calculate the examinee's summed score, X . In practice, we would have a `tables` command for each item to be investigated. However, because we are examining if item 3 is exhibiting DIF we have a single `tables` command. The `tables` command instructs SAS to create a three-way contingency table of summed score by examinee race by item 3's responses (i.e., `tables x*race*i3`). The subcommand `CMH` on the `tables` line indicates the calculation of the MH statistic and `NOPRINT` suppresses printing the 2×2 table series; `CMH` is the initialization for the Cochran–Mantel–Haenszel statistic.

The abridged output is presented in Table 12.2. At the 5% significance level, our MH value of 383.1077 (line labeled Nonzero Correlation) is significant and reflects the expected DIF; $df = 1$.⁵ (The actual value of MH χ^2 calculated according to Equation 12.1 is 381.0889, with the difference reflecting the use of the “0.5” correction in Equation 12.1, but not with the `CMH` statistic)

Given our coding, we take the reciprocal of the value on the Odds Ratio Mantel–Haenszel line (i.e., 12.2054) to obtain $\hat{\alpha}_{MH} = OR = 0.0819$, with the magnitude of item 3's DIF being $\hat{\beta}_{MH} = \ln(0.0819) = -2.5019$.⁶ (Endnote 7 shows the results using SPSS and SYSTAT (SYSTAT, 2017).) Recall that if α_{MH} is less than 1, then the item favors the Focal group. Conversely, an α_{MH} greater than 1 indicates that the item favors the Reference group with an α_{MH} of 1 indicating neutrality. For our example item, our $\hat{\alpha}_{MH}$ of 0.0819 indicates that this item favors the Focal group (i.e., African Americans). Therefore, (1) given the significant MH statistic, this item is exhibiting significant DIF across the two groups, (2) the odds of African Americans (i.e., the Focal group) correctly responding to this item are, on average, more than 12 times (i.e., $1/0.0819$) the odds of Caucasians of similar proficiency (i.e., the Reference group), and (3) on average and in terms of the log odds of success, African Americans find this item to be about two and a half times easier than Caucasians of comparable proficiency do (i.e., $\hat{\beta}_{MH} = -2.5019$). The significant MH would be interpreted as indicating the presence of uniform DIF on item 3.

Although for our example we focus on item 3, in practice, we would have obtained MH χ^2 , $\hat{\alpha}_{MH}$, and $\hat{\beta}_{MH}$ for each item. Subsequently, any item that exhibits significant and meaningful DIF would be removed, and the MH χ^2 , $\hat{\alpha}_{MH}$, and $\hat{\beta}_{MH}$ for each remaining item would be recalculated. For instance, in our example we would remove item 3 given the magnitude of its $\hat{\beta}_{MH}$ and its significant MH χ^2 , recalculate our X , and then examine

TABLE 12.2. Abridged MH Output

The FREQ Procedure

**Summary Statistics for race by i3
Controlling for X**

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	383.1077	<.0001
2	Row Mean Scores Differ	1	383.1077	<.0001
3	General Association	1	383.1077	<.0001

Common Odds Ratio and Relative Risks

Statistic	Method	Value	95% Confidence Limits	
Odds Ratio	Mantel-Haenszel	12.2054	9.2036	16.1863
	Logit **	11.1333	8.4504	14.6680
Relative Risk (Column 1)	Mantel-Haenszel	4.7866	3.9720	5.7683
	Logit **	2.8010	2.4462	3.2072
Relative Risk (Column 2)	Mantel-Haenszel	0.5099	0.4726	0.5502
	Logit **	0.6355	0.6014	0.6715

** These logit estimators use a correction of 0.5 in every cell of those tables that contain a zero. Tables with a zero row or a zero column are not included in computing the logit estimators.

Breslow-Day Test for Homogeneity of the Odds Ratios	
Chi-Square	8.9482
DF	8
Pr > ChiSq	0.3467

Total Sample Size = 2030

the remaining items for evidence of DIF. We would repeat the process until we are satisfied that the evidence does not indicate items are exhibiting meaningful DIF.

As indicated above, whether item 3 is biased against Reference group members is not determined by the above statistical analysis. Rather, item 3 would be subjected to panel review to determine whether it is biased and appropriate corrective actions should be taken. (One may conjecture that the differential performance on this item is due to an opportunity to learn the relevant material.)

Example: DIF Analysis of Vocabulary Test, `mantelhaen.test` and `difR`

Base R has a function for calculating the Cochran–Mantel–Haenszel statistic. We first use this built-in function, `mantelhaen.test`, and then repeat the analysis using the `difR` package. Our R session is presented in Table 12.3. As we did in previous chapters, we verify the data are correctly read using the `head` and `tail` functions. Subsequently, we remove the `id` variable and obtain the frequency distributions for `race` and each item using the `table` function. (Rather than repeatedly call the `table` function for `race` and each item, we use the `apply` function.) Next, we calculate the X for our 10 items and examine its frequency distribution (`table(X)`) for anomalies.

A basic approach for obtaining the MH value for item 3 is to first create its 2×2 table series conditional on X using the `table` function (`itm3 = table(vocabtbl$race, vocabtbl$i3, X)`) and then pass this three-dimensional `itm3` table object as an argument to the `mantelhaen.test` function (`mantelhaen.test(itm3)`). However, we combine these two steps into one (`itm3MH = mantelhaen.test((itm3C = with(vocabtbl, table(race,i3,X))))`) while saving the results to the output object `itm3MH`. By default, the `mantelhaen.test` function uses the continuity correction. Consequently, item 3's MH χ^2 matches our hand calculation (i.e., 381.09). As is the case with the SAS analysis, we take the reciprocal of the common odds ratio to obtain $\hat{\alpha}_{MH} = 0.0819$ (`alphaMHitm3 = 1/itm3MH$estimate`) with an effect size of $\hat{\beta}_{MH} = \ln(\hat{\alpha}_{MH}) = -2.5019$. In short, this B item is correctly identified as exhibiting DIF. The interpretation from our SAS analysis applies here. To obtain the 2×2 table series analyzed by the MH we call the `Desc` function passing to it the table object `itm3`; this package also contains a `BreslowDayTest` function. Our 2×2 table series have the X as the columns, with the item responses (`i3`) nested within `race` as the rows of the table. As an example, for $X = 2$ we have

	X	
<code>race</code>	0	1
0	7	0
1	24	18
Sum	31	18

In practice, we would repeat the above `mantelhaen.test` process for each item; a large set of items using a looping command (e.g., a `for` loop) would be convenient. Alternatively, we can use the `difR` package. This package contains a plethora of DIF IRT- and nonIRT-based techniques. For our purposes, we use the `difMH` function (Table 12.4). Our `difMH` call specifies the data matrix (`vocabtbl`), the grouping variable (`group = "race"`), and the coding for the focal group (`focal.name = 1`), and saves the function's output to `vocabMH`.

The Mantel–Haenszel Chi-square statistic table shows that all items except item 6 exhibit significant DIF, with seven items classified in ETS's C category. Figure 12.3 shows the items MH χ^2 s. The Items detected as DIF items section simply tells us which items had significant MH χ^2 s. The subsequent table, Effect size

TABLE 12.3. Base R Session for DIF Analysis—MH

```

> library(DescTools) # for Desc function
> packageVersion("DescTools")
[1] '0.99.37'

> vocab=read.table("Vocab.dat",col.names=c("id","race",paste0("i",1:10)))

> head(vocab,5)
   id race i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
1 1 0 1 1 1 0 1 1 1 1 1 0
2 2 0 0 1 1 1 1 0 0 1 0 1
3 3 0 1 1 0 0 1 1 0 1 0 1
4 4 0 1 0 0 0 1 1 0 1 0 1
5 5 0 1 1 1 1 1 0 1 1 1

> tail(vocab,5)
   id race i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
2026 2026 1 0 0 1 0 1 0 0 0 0 0
2027 2027 1 1 0 0 0 1 1 0 1 0 0
2028 2028 1 0 0 1 1 1 0 0 1 1 0
2029 2029 1 0 1 0 1 0 1 0 0 0 1
2030 2030 1 1 0 1 0 1 0 0 1 0 0

> vocabtbl=within(vocab,rm(id))

> apply(vocabtbl,2,table)
   race i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
0 1022 552 756 664 923 204 502 1816 243 1557 768
1 1008 1478 1274 1366 1107 1826 1528 214 1787 473 1262

> X = rowSums(vocabtbl[, c(paste0("i",1:10))]) # calculate summed score

> table(X)
X
 1 2 3 4 5 6 7 8 9 10
9 49 116 233 347 387 403 340 130 16

> # race = 0 (reference), race = 1 (focal)
> perform MH on item 3, save MH results to itm3MH, & create 2 x 2 x K tables
  for item 3
> print((itm3MH=mantelhaen.test((itm3C=with(vocabtbl,table(race,i3,X))))))

  Mantel-Haenszel chi-squared test with continuity correction

  data: (itm3C = with(vocabtbl, table(race, i3, X)))
  Mantel-Haenszel X-squared = 381.09, df = 1, p-value < 2.2e-16
  alternative hypothesis: true common odds ratio is not equal to 1
  95 percent confidence interval:
  9.20357 16.18631
  sample estimates:
  common odds ratio
  12.2054

```

(continued)

TABLE 12.3. (continued)

```

> print((alphaMHitm3=1/itm3MH$estimate))
  common odds ratio
  0.08193094

> print((betaaMHitm3=log(alphaMHitm3)))
  common odds ratio
  -2.501879

> Desc(itm3)
-----
itm3 (table)

Summary:
n: 2'030, 3-dim table: 2 x 2 x 10

Chi-squared test for independence of all factors:
X-squared = 689.569, df = 28, p-value = < 2.2e-16

Warning message:
  Exp. counts < 5: Chi-squared approx. may be incorrect!!

      X    1    2    3    4    5    6    7    8    9    10   Sum
race i3
0    0    1    7   20   46   78  111  127   97   16    0  503
      1    0    0    8   22   31   88  127  151   79   13  519
1    0    6   24   26   41   35   16   10    3    0    0  161
      1    2   18   62  124  203  172  139   89   35    3  847
Sum  0    7   31   46   87  113  127  137  100   16    0  664
      1    2   18   70  146  234  260  266  240  114   16 1366

```

(ETS Delta scale), provides us with our $\hat{\alpha}_{MH}$ (`alphaMH`) and D_{MH} (`deltaMH`) with $\hat{\beta}_{MH} = D_{MH}/-2.35$. As can be seen, for item 3 $\hat{\alpha}_{MH} = 0.0819$ and $D_{MH} = 5.8794$ ($\hat{\beta}_{MH} = -2.502$). Given item 3's effect size ($\hat{\beta}_{MH}$ or D_{MH}), it may contribute to distorting the other items' MH χ^2 s. Accordingly, we remove item 3 and recalculate the remaining items' MH χ^2 s. Iteration 2's results continue to show significant MH χ^2 s for all items (except item 4), albeit with fewer items classified in the C category than in the first iteration. We remove item 7 given the magnitude of its effect size; item 7 is the second B item on the test. Our third iteration's results show only items 5 and 9 as C category items, with item 9 having a larger D_{MH} than item 5. Therefore, in iteration 4 we remove item 9, examine the results, and then remove item 6. The results from our fifth iteration show that although item 4 has a significant MH χ^2 , it is classified in category B. Our remaining six items are not exhibiting meaningful DIF. As mentioned above, items that are flagged for DIF are examined to determine if they are biased.

As mentioned above, a negative β_{MH} (or a positive D_{MH}) indicates that, on average, Focal group members tended to provide a correct response more often than comparable Reference group members. Moreover, the Reference group IRF is to the right of the Focal group IRF by an amount "equal" to the magnitude of the DIF and reflecting that the item

TABLE 12.4. difR Analysis—MH

```

> # This is a continuation of the session from Table 12.3

> library(difR)
> packageVersion("difR")
[1] '5.1'

> # iteration 1
> print((vocabMH=difMH(vocabtbl, group="race", focal.name=1)))

Detection of Differential Item Functioning using Mantel-Haenszel method
with continuity correction and without item purification

Results based on asymptotic inference

Matching variable: test score

No set of anchor items was provided

No p-value adjustment for multiple comparisons

Mantel-Haenszel Chi-square statistic:

      Stat.    P-value
i1   37.8137  0.0000 ***
i2   49.0270  0.0000 ***
i3   381.0889 0.0000 ***
i4   19.0730  0.0000 ***
i5   31.9941  0.0000 ***
i6   0.3712   0.5424
i7   61.8708  0.0000 ***
i8   25.3682  0.0000 ***
i9   9.1980   0.0024 **
i10  41.9680  0.0000 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Detection threshold: 3.8415 (significance level: 0.05)

Items detected as DIF items:

i1
i2
i3
i4
i5
i7
i8
i9
i10

Effect size (ETS Delta scale):

Effect size code:
'A': negligible effect
'B': moderate effect
'C': large effect

```

(continued)

TABLE 12.4. (*continued*)

```

alphaMH deltaMH
i1  2.1032 -1.7471 C
i2  2.2357 -1.8907 C
i3  0.0819  5.8794 C
i4  1.6218 -1.1363 B
i5  3.3523 -2.8426 C
i6  0.9183  0.2002 A
i7  0.2489  3.2677 C
i8  2.6439 -2.2848 C
i9  0.6769  0.9171 A
i10 2.1330 -1.7802 C

Effect size codes: 0 'A' 1.0 'B' 1.5 'C'
(for absolute values of 'deltaMH')

Output was not captured!

> plot(vocabMH)          # produces Figure 12.3

> # iteration 2
> vocabtblpurfctn=within(vocabtbl,rm(i3))

> difMH(vocabtblpurfctn, group="race", focal.name=1)

:

Mantel-Haenszel Chi-square statistic:

      Stat.   P-value
i1  11.1139  0.0009 ***
i2  11.8468  0.0006 ***
i4  0.6676  0.4139
i5  15.1094  0.0001 ***
i6  11.9953  0.0005 ***
i7  89.5825  0.0000 ***
i8  9.4039  0.0022 **
i9  36.3174  0.0000 ***
i10 8.4537  0.0036 **

:
Effect size (ETS Delta scale):

Effect size code:
'A': negligible effect
'B': moderate effect
'C': large effect

alphaMH deltaMH
i1  1.5352 -1.0074 B
i2  1.5152 -0.9766 A
i4  1.1066 -0.2381 A
i5  2.3283 -1.9860 C
i6  0.6229  1.1125 B
i7  0.1691  4.1766 C
i8  1.8627 -1.4618 Ba

```

(continued)

TABLE 12.4. (continued)

```

i9   0.4419  1.9191 C
i10  1.4346 -0.8481 A

:
> # iteration 3
> vocabtblpurfctn=within(vocabtblpurfctn,rm(i7))

> difMH(vocabtblpurfctn, group="race", focal.name=1)

:

Mantel-Haenszel Chi-square statistic:

      Stat.    P-value
i1    6.9021  0.0086 **
i2    5.9822  0.0145 *
i4    0.0320  0.8579
i5   11.5675  0.0007 ***
i6   17.6224  0.0000 ***
i8   6.8363  0.0089 **
i9   63.0169  0.0000 ***
i10  3.4316  0.0640 .

:
Effect size (ETS Delta scale):

Effect size code:
'A': negligible effect
'B': moderate effect
'C': large effect

      alphaMH deltaMH
i1    1.4119 -0.8106 A
i2    1.3562 -0.7160 A
i4    0.9721  0.0665 A
i5    2.1277 -1.7744 C
i6    0.5564  1.3778 B
i8    1.7259 -1.2825 B
i9    0.2859  2.9422 C
i10   1.2671 -0.5564 A
:

> # iteration 4
> vocabtblpurfctn=within(vocabtblpurfctn,rm(i9))
> difMH(vocabtblpurfctn, group="race", focal.name=1)
:
Mantel-Haenszel Chi-square statistic:

      Stat.    P-value
i1    3.2073  0.0733 .
i2    1.2713  0.2595
i4    2.2568  0.1330
i5    6.8178  0.0090 **

```

(continued)

TABLE 12.4. (*continued*)

```

i6 27.4317 0.0000 ***
i8 3.6884 0.0548 .
i10 0.2365 0.6267
:
Effect size (ETS Delta scale):

Effect size code:
'A': negligible effect
'B': moderate effect
'C': large effect

alphaMH deltaMH
i1 1.2748 -0.5706 A
i2 1.1626 -0.3541 A
i4 0.8264 0.4482 A
i5 1.8542 -1.4511 B
i6 0.4741 1.7541 C
i8 1.5109 -0.9698 A
i10 1.0730 -0.1656 A
:

> # iteration 5
> vocabtblpurfctn=within(vocabtblpurfctn,rm(i6))
> difMH(vocabtblpurfctn, group="race", focal.name=1)
:
Mantel-Haenszel Chi-square statistic:

      Stat.  P-value
i1 1.3028 0.2537
i2 0.0678 0.7945
i4 6.8126 0.0091 **
i5 3.7199 0.0538 .
i8 1.3625 0.2431
i10 0.5041 0.4777

:
Effect size (ETS Delta scale):

Effect size code:
'A': negligible effect
'B': moderate effect
'C': large effect

alphaMH deltaMH
i1 1.1729 -0.3748 A
i2 1.0431 -0.0992 A
i4 0.7142 0.7909 A
i5 1.5973 -1.1005 B
i8 1.3007 -0.6178 A
i10 0.9015 0.2436 A
:

```

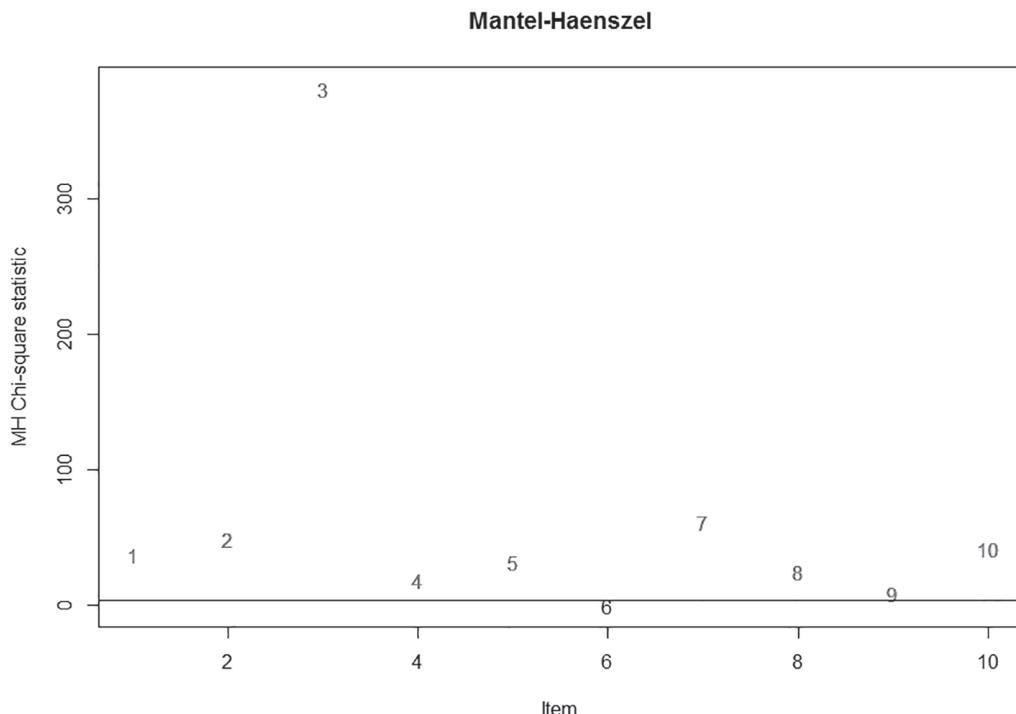
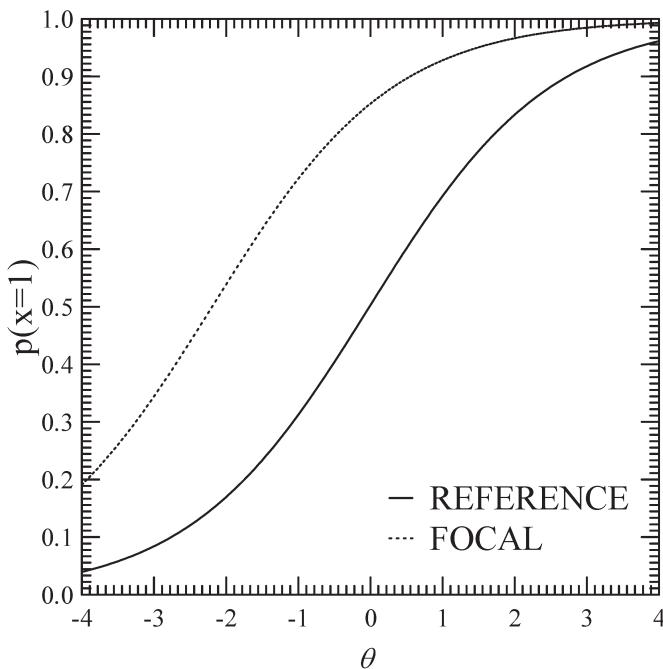


FIGURE 12.3. MH χ^2 values. All items above the horizontal line are significant values; the numbers displayed are ordinal positions.

is more difficult for the Reference group. As a demonstration, we perform a 1PL model calibration first for the Reference group and then for the Focal group using SYSTAT. We use the mean–mean linking approach to align the Focal group metric with that of the Reference group. After linking our metrics, item 3 has an estimated location of -2.1946 for the Focal group, whereas for the Reference group this item is estimated to be located at -0.0139. Clearly, the item is more difficult for the Reference group than for the Focal group. This is reflected in the corresponding IRFs (Figure 12.4).

Example: DIF Analysis of Vocabulary Test, SAS proc logistic

There are several ways to perform a logistic regression in SAS (e.g., `proc logistic`, `proc genmod`, `proc catmod`). For our analysis we use `proc logistic`. Table 12.5 contains the SAS program for performing the various model comparisons. As with the MH analysis, the line `X = sum (of i1-i10)` creates the summed score, X. For this analysis, we re-code the Reference group to be 1 and the Focal group to be 0. Consequently, we do not have to take the reciprocal of the outputted odds ratio to obtain the appropriate odds ratio (OR). Because by default SAS models the outcome variable's 0 value, we use the descending option to have SAS predict a response of 1 on our item.

**FIGURE 12.4.** IRFs for item 3.**TABLE 12.5. SAS Program for Performing Logistic Regression DIF Procedure on Item 3**

```

ods graphics on;
title "DIF analysis, item 3, reference grp:1, focal grp:0. ";
data d1;
  infile "C:\vocab.dat";
  input person raceR0 i1-i10;
  X=sum (of i1-i10);
  raceR1=0;
  if (raceR0=0) then raceR1=1;
run;

proc logistic descending; /* full model */
  title "Full model";
  model i3 = x raceR1 x*raceR1/rsquare;
run;

proc logistic descending; /* reduce model (1)/*full' model */;
  title "reduce model (1)/*full' model";
  model i3 = x raceR1/rsquare;
  oddsratio raceR1;
run;

proc logistic descending; /* reduce model (2) */;
  title "reduce model (2)";
  model i3 = x/rsquare;
  oddsratio x ;
run;

```

The first step is to compare the full model with the reduced model (1) to investigate the existence of nonuniform DIF. These models are

$$\text{full model: } \hat{z}_j = \beta_0 + \beta_1(X) + \beta_2 * \text{raceR1} + \beta_3(X * \text{raceR1})$$

and

$$\text{reduced model (1)/ 'full' model: } \hat{z}_j = \beta_0 + \beta_1(X) + \beta_2 * \text{raceR1}.$$

The comparison of these two models determines whether the interaction term ($X * \text{raceR1}$) significantly improves model fit. If there is a significant improvement, then there is evidence of nonuniform DIF. Accordingly, the item of interest, $i3$, is regressed on the predictors X , race, and the interaction of X and race in the full model (i.e., model $i3 = x \text{ raceR1 } x * \text{raceR1}$). The subsequent logistic regression is the reduced model (1)/'full' model: (i.e., model $i3 = x \text{ raceR1}$). Depending on the outcome of comparing these two regressions, a second analysis is performed. This analysis determines whether the inclusion of raceR1 as a predictor leads to a significant improvement in model fit over simply using the summed score. Stated another way, we are comparing the uniform DIF model to the no DIF model. The corresponding models are

$$\text{reduced model (1)/ 'full' model: } \hat{z}_j = \beta_0 + \beta_1(X) + \beta_2 * \text{raceR1}$$

and

$$\text{reduced model (2): } \hat{z}_j = \beta_0 + \beta_1(X).$$

If there is a significant difference between these two models, then we know that the race predictor accounts for a significant improvement in model fit to item j . In other words, there is evidence of uniform DIF. The corresponding output is presented in Table 12.6.

Convergence is achieved for all analyses. From the Model Fit Statistics section, the last entry in the Intercept and Covariates column labeled -2 Log L provides us with the models' log likelihood statistics. Specifically,

$$\text{full model: } -2 \ln L_F = 2084.631$$

and for the

$$\text{reduced model (1)/full' model: } -2 \ln L_{f/R1} = 2088.655.$$

For determining the presence of nonuniform DIF, we have

$$\Delta G^2 = (-2 \ln L_{f/R1}) - (-2 \ln L_F) = 2088.655 - 2084.631 = 4.024.$$

**TABLE 12.6. Abridged Logistic Regression Output:
Full Model and Reduced Model (1)**

Full model

The LOGISTIC Procedure

:

Response Profile		
Ordered Value	i3	Total Frequency
1	1	1366
2	0	664

Probability modeled is i3=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2568.331	2092.631
SC	2573.947	2115.094
-2 Log L	2566.331	2084.631

R-Square 0.2112 Max-rescaled R-Square 0.2944

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	481.7002	3	<.0001
Score	431.4481	3	<.0001
Wald	325.4422	3	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1767	0.2826	17.3400	<.0001
X	1	0.5830	0.0605	92.8958	<.0001
raceR1	1	-1.6994	0.4242	16.0464	<.0001
X*raceR1	1	-0.1509	0.0758	3.9595	0.0466

:

reduce model (1) /'full' model - monotonic decreasing predicted p functions

The LOGISTIC Procedure

:

Probability modeled is i3=1.

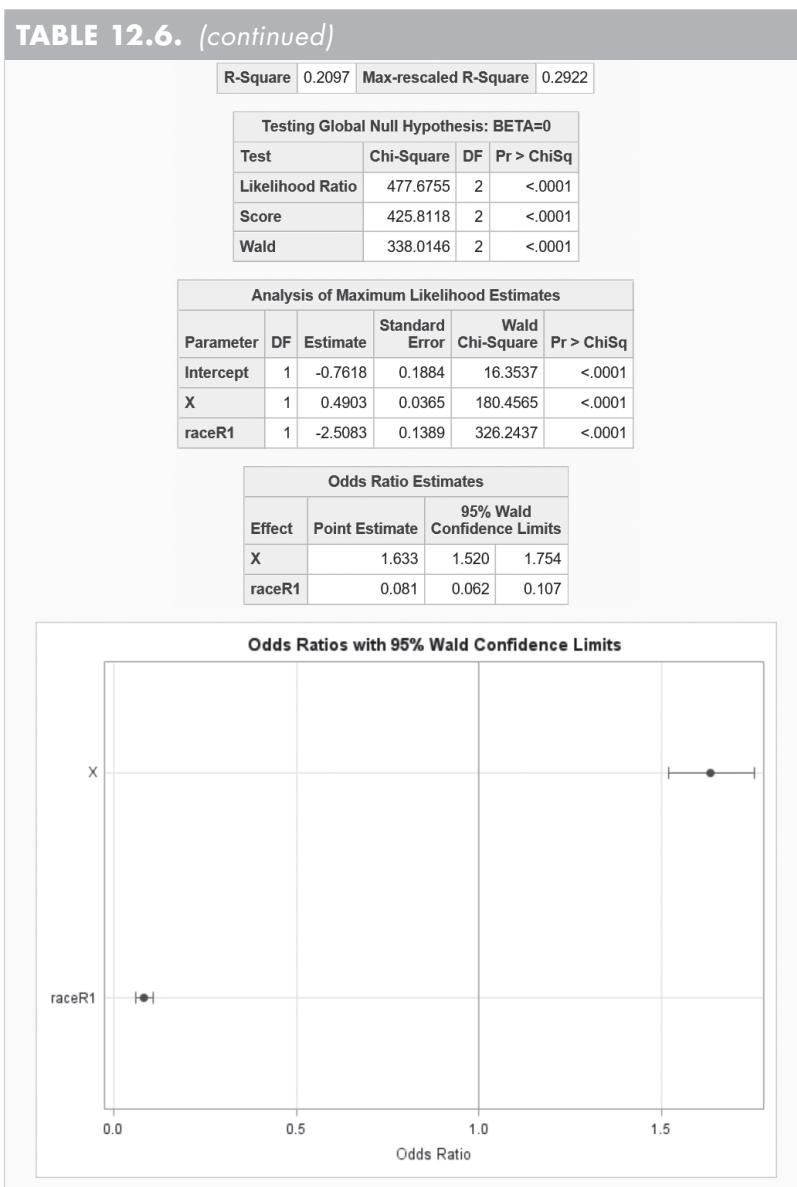
Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2568.331	2094.655
SC	2573.947	2111.503
-2 Log L	2566.331	2088.655

(continued)

TABLE 12.6. (continued)

With 1 *df* the *p* of observing $\Delta G^2 = 4.024$ when the race by *X* interaction is 0 (i.e., the null hypothesis is true) is 0.0449. As a result, at the 5% significance level, we have a significant likelihood ratio test (i.e., there is evidence of nonuniform DIF). To obtain a measure of the effect size, we examine the difference in R^2 s. SAS provides both the Cox and Snell R^2 (labeled R-Square) as well as the Nagelkerke R^2 (labeled Max-rescaled R-Square). Using the Nagelkerke R^2 s for this item, we have an $R_F^2 = 0.2944$ for the full model, and for the reduced model (1) $R_{R1}^2 = 0.2922$. The effect of our nonuniform DIF is

$$\Delta R^2 = R_F^2 - R_{R1}^2 = 0.2944 - 0.2922 = 0.0022.$$

Therefore, although this item is exhibiting significant nonuniform DIF, ΔR^2 indicates it is exhibiting negligible DIF.

Using the values from the Analysis of Maximum Likelihood Estimates tables, our estimated models are

$$\begin{aligned}\text{full model: } \hat{z}_3 &= b_0 + b_1(X) + b_2 * \text{raceR1} + b_3(X * \text{raceR1}) \\ &= -1.1767 + 0.5830(X) - 1.6994 * \text{raceR1} - 0.1509(X * \text{raceR1})\end{aligned}$$

and

$$\begin{aligned}\text{reduced model (1)}/\text{'full' model: } \hat{z}_3 &= b_0 + b_1(X) + b_2 * \text{raceR1} \\ &= -0.7618 + 0.4903(X) - 2.5083 * \text{raceR1}.\end{aligned}$$

Normally, when we find significant nonuniform DIF, we would not proceed to examine the item for uniform DIF.⁸ This is tantamount to obtaining a significant interaction term in ANOVA and proceeding to interpret the main effects. However, because the magnitude of the nonuniform DIF is small, and for pedagogical reasons, we proceed to perform a second analysis for uniform DIF. Table 12.7 contains the results of this analysis. For the reduced model (2), we have a $-2\ln L_{R2} = 2525.103$, and from above we have a $-2\ln L_{f/R1} = 2088.655$ for the uniform DIF model (i.e., reduced model (1)/'full' model). Therefore, comparing the reduced model (2) with the reduced model (1)/'full' model, we have

$$\Delta G^2 = (-2\ln L_{R2}) - (-2\ln L_{f/R1}) = 2525.103 - 2088.655 = 436.448.$$

With 1 *df* this ΔG^2 is significant at the 5% significance level and indicates the presence of uniform DIF. (This test is similar to the MH test performed above [i.e., a test for uniform DIF].) Stated another way, the inclusion of the `raceR1` predictor leads to a significant improvement in fit vis à vis its absence (i.e., reduced model (2)). The corresponding effect size ΔR^2 for this significant ΔG^2 is:

$$\Delta R^2 = R_{R1}^2 - R_{R2}^2 = 0.2922 - 0.0280 = 0.2642.$$

According to Jodoin and Gierl's (2001) guidelines, this is a C-level (large DIF) item.

As mentioned above, β_2 indicates the difference between the Reference and Focal groups' average performance (in terms of the log odds of success) on the item of interest (i.e., the degree of DIF), with β_2 's sign indicating the directionality of the DIF. Given our coding of the `raceR1` variable, if $\beta_2 < 0$, then the item favors the Focal group; if $\beta_2 > 0$, then the item favors the Reference group.⁹ As can be seen, our b_2 approximately equals our $\hat{\beta}_{MH}$ from our MH analysis. Graphically (Table 12.6), the Odds Ratio with a confidence limits plot shows that $OR = 0.081$, for `raceR1` is clearly significantly different than the no DIF reference line ($OR = 1.0$).

TABLE 12.7. Abridged Logistic Regression Output for Reduced Model (2)

reduce model (2)					
The LOGISTIC Procedure					
:					
Probability modeled is i3=1.					
Model Convergence Status					
Convergence criterion (GCONV=1E-8) satisfied.					
Model Fit Statistics					
Criterion	Intercept Only	Intercept and Covariates			
AIC	2568.331	2529.103			
SC	2573.947	2540.335			
-2 Log L	2566.331	2525.103			
R-Square	0.0201	Max-rescaled R-Square 0.0280			
Testing Global Null Hypothesis: BETA=0					
Test	Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio	41.2278	1	<.0001		
Score	41.2509	1	<.0001		
Wald	40.4863	1	<.0001		
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2963	0.1647	3.2351	0.0721
X	1	0.1703	0.0268	40.4863	<.0001
Odds Ratio Estimates					
Effect	Point Estimate		95% Wald Confidence Limits		
X	1.186		1.125	1.249	

To summarize, the MH statistic identified item 3 as exhibiting DIF that would typically be interpreted as uniform DIF. The logistic regression analysis indicated significant nonuniform DIF (albeit with a small effect size) as well as significant uniform DIF. To explain why the MH statistic is able to identify an item exhibiting nonuniform DIF, we perform the simple logistic regression of the item on the summed score separately for each manifest group. The results for the Reference and Focal groups are presented in Table 12.8. The resulting models are

$$\text{Reference group} \quad \hat{z}_3 = b_{0R} + b_{1R}X = -2.8760 + 0.4321(X)$$

and

$$\text{Focal group} \quad \hat{z}_3 = b_{0F} + b_{1F}X = -1.1768 + 0.5830(X).$$

A comparison of the two models' intercept estimates (i.e., b_{0F} and b_{0R}) shows they are unequal and are not within 2 standard errors of one another. Similarly, comparing the two models' regression coefficient estimates (i.e., b_{1F} and b_{1R}), we see they are unequal and are not within one standard error of each other. These differences reflect the DIF of the item. We can relate these differences to the full model's estimates from Table 12.6. For instance, the difference between b_{0F} and b_{0R} equals the coefficient for the raceR1 variable in the full model

$$b_2 = b_{0R} - b_{0F} = -2.8760 - (-1.1768) = -1.6992$$

It is this difference in the intercepts that the MH χ^2 is, in effect, capturing. With respect to the regression coefficients, the difference between b_{1F} and b_{1R} equals the coefficient for the raceR1 by X interaction term in the full model (Table 12.6)

$$b_3 = b_{1R} - b_{1F} = 0.4321 - 0.5830 = -0.1509.$$

The magnitude of b_2 and b_3 indicate the degree of uniform and nonuniform DIF, respectively.

Figure 12.5 shows the logit regression lines for each model. As can be seen, the Focal group members have an advantage over comparable Reference group members throughout the summed score metric. However, we can see that this advantage varies slightly as a function of the summed score and thereby reflects nonuniform DIF. In this case, there is an ordinal interaction present with this item. This slight lack of parallelism is reflective of the significant interaction term with a small effect size. Swaminathan and Rogers (1990) conjecture that in the case of nonuniform DIF, the MH statistic may do a better job of detecting an ordinal interaction than a disordinal interaction. (A disordinal interaction occurs when the logit regression lines cross within the summed score range.) Given the possibility of cancellation, the MH statistic is not designed for detecting non-uniform DIF. However, as this example shows, the MH statistic may detect some cases of nonuniform DIF.

As done above with the MH χ^2 approach, in practice we would continue our DIF analysis with the remaining items. Moreover, any item identified as having significant and meaningful DIF would be subjected to a panel review to determine whether the item is bias.

Example: DIF Analysis of Vocabulary Test, `glm` and `difR`

Parallel to our MH analysis in R, we begin by using the Base system's generalized linear model (`glm`) function to perform a logistic regression analysis of item 3, followed

TABLE 12.8. Abridged Logistic Regression Output for Reference and Focal Groups**Reference Group Results**

:

Number of Observations Read	1022
Number of Observations Used	1022

Response Profile		
Ordered Value	i3	Total Frequency
1	1	519
2	0	503

Probability modeled is i3=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1418.542	1317.517
SC	1423.472	1327.376
-2 Log L	1416.542	1313.517

R-Square 0.0959 Max-rescaled R-Square 0.1279

:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8760	0.3164	82.6110	<.0001
X	1	0.4321	0.0457	89.2355	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
X	1.540	1.408	1.685

:

(continued)

TABLE 12.8. (*continued*)**Focal Group Results**

:

Number of Observations Read	1008
Number of Observations Used	1008

Response Profile		
Ordered Value	i3	Total Frequency
1	1	847
2	0	161

Probability modeled is i3=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	887.445	775.114
SC	892.361	784.945
-2 Log L	885.445	771.114

R-Square	0.1072	Max-rescaled R-Square	0.1834
----------	--------	-----------------------	--------

:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1768	0.2826	17.3433	<.0001
X	1	0.5830	0.0605	92.9019	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
X	1.791	1.591	2.017

:

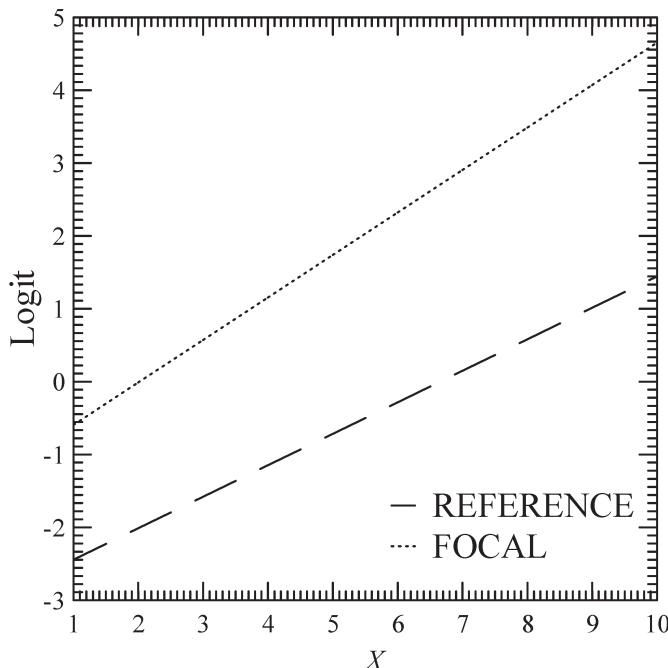


FIGURE 12.5. Logit regression lines for Reference and Focal groups.

by using `difR`. Table 12.9 contains our R session. After reading our data, calculating the summed score X , and removing the case `id` variable, we re-code our manifest groups, so the Reference group is coded 1 and the Focal group is coded 0, as we did in our SAS logistic regression analysis. Using the `table` function, we verify that the re-coded race variable (`raceR1`) is correct (e.g., the frequency for `raceR1 = 1` matches the `raceR0 = 0` count).

We begin our DIF analysis by estimating the Full model (`glm(vocab$i3 ~ X + vocab$raceR1 + X*vocab$raceR1, . . .)`), followed by the reduced model (1)/*full' model (`glm(vocab$i3 ~ X + vocab$raceR1, . . .)`) and the reduced model (2) (`glm(vocab$i3 ~ X , . . .)`). In our calls to `glm`, we specify the logit link function and the binomial distribution family. By default, the `glm` function performs up to 25 iterations unless otherwise changed (e.g., `glm(vocab$i3 ~ X . . . , control = list(maxit = 50))`). As can be seen, each of our models converged in either four or five iterations (Number of Fisher Scoring iterations).

We use the `anova` function to obtain our model fit and comparison statistics. From the `Terms added sequentially (first to last)` table, we obtain our G^2 s for the full model, reduced model (1)/*full' model, and reduced model (2). The `Resid. Dev` column shows the deviance statistic for the full model as 2084.6 ($X:vocab$raceR1$ line), and for the reduced model (1)/*full' model we have $(-2\ln L_{fR1}) = 2088.7$ ($vocab$raceR1$ line) with the $\Delta G^2 = 4.02$ (Deviance column). From above, we know this ΔG^2 is significant ($p = 0.0449$) although it is associated with negligible DIF. Similarly, the comparison between the reduced model (1)/*full' model

TABLE 12.9. Base R Session for DIF Analysis—Logistic Regression

```

> library(car)                                     # for recode function
> packageVersion("car")
[1] '3.0.3'

> library(DescTools)                            # for Desc & PseudoR2 function
> packageVersion("DescTools")
[1] '0.99.37'

> vocab=read.table("Vocab.dat",col.names=c("id","raceR0",paste0("i",1:10)))

> head(vocab,5)
   id raceR0 i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
1  1      0  1  1  1  0  1  1  1  1  1  0
2  2      0  0  1  1  1  1  0  0  1  0  1
3  3      0  1  1  0  0  1  1  0  1  0  1
4  4      0  1  0  0  0  1  1  0  1  0  1
5  5      0  1  1  1  1  1  1  0  1  1  1

> X = rowSums(vocab[, c(paste0("i",1:10))])      # calculate summed score

> vocab=within(vocab,rm(id))                      # remove case label

> vocab$raceR1=recode(vocab$raceR0,"1=0;0=1")    # recode manifest groups, Ref = 1

> apply(vocab,2,table)
   raceR0   i1   i2   i3   i4   i5   i6   i7   i8   i9   i10 raceR1
0    1022  552  756  664  923  204  502 1816  243 1557  768  1008
1    1008 1478 1274 1366 1107 1826 1528  214 1787  473 1262  1022

> Desc(X)
-----
X (numeric)

  length     n    NAs unique    0s    mean  meanCI'
  2'030  2'030      0     10      0    6.07    5.99
  100.0% 100.0%      0.0%      0.0%           6.14

  .05     .10     .25 median    .75    .90    .95
  3.00    4.00    5.00      6.00    7.00    8.00   9.00

  range      sd vcoef      mad    IQR    skew    kurt
  9.00    1.79   0.30     1.48    2.00   -0.28  -0.47

  level freq   perc cumfreq cumperc
  1      1     9  0.4%      9   0.4%
  2      2    49  2.4%     58   2.9%
  3      3   116  5.7%    174   8.6%
  4      4   233 11.5%    407  20.0%
  5      5   347 17.1%    754  37.1%
  6      6   387 19.1%   1'141  56.2%
  7      7   403 19.9%   1'544  76.1%
  8      8   340 16.7%   1'884  92.8%
  9      9   130  6.4%   2'014 99.2%
 10     10    16  0.8%   2'030 100.0%

' 95%-CI (classic)
```

(continued)

TABLE 12.9. (continued)

```

> # Full model
> full_mod = glm(vocab$i3 ~ X + vocab$raceR1 + X*vocab$raceR1, family=
+ binomial(link="logit"))

> anova(full_mod)
Analysis of Deviance Table

Model: binomial, link: logit

Response: vocab$i3

Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev
NULL           2029      2566.3
X              1     41.23    2028      2525.1
vocab$raceR1   1    436.45    2027      2088.7
X:vocab$raceR1 1      4.02    2026      2084.6

> summary(full_mod)
Call:
glm(formula = vocab$i3 ~ X + vocab$raceR1 + X * vocab$raceR1,
family = binomial(link = "logit"))

Deviance Residuals:
      Min        1Q     Median       3Q       Max
-2.6524 -0.9378  0.4327  0.9427  1.8800

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17680  0.28258 -4.165 3.12e-05 ***
X             0.58303  0.06049  9.639 < 2e-16 ***
vocab$raceR1 -1.69926  0.42424 -4.005 6.19e-05 ***
X:vocab$raceR1 -0.15094  0.07584 -1.990  0.0466 *
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2566.3 on 2029 degrees of freedom
Residual deviance: 2084.6 on 2026 degrees of freedom
AIC: 2092.6

Number of Fisher Scoring iterations: 5

> PseudoR2(full_mod, which = c("CoxSnell", "Nagelkerke"))
CoxSnell Nagelkerke
0.2112381 0.2943939

> # reduced model (1) / 'full'
> red_mod1 = glm(vocab$i3 ~ X + vocab$raceR1, family=binomial(link="logit"))

> summary(red_mod1)
Call:
glm(formula = vocab$i3 ~ X + vocab$raceR1, family = binomial(link = "logit"))

Deviance Residuals:
      Min        1Q     Median       3Q       Max
-2.5306 -1.0415  0.4628  0.8946  1.9760

```

(continued)

TABLE 12.9. (*continued*)

```

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.76175 0.18836 -4.044 5.25e-05 ***
X 0.49026 0.03649 13.434 < 2e-16 ***
vocab$raceR1 -2.50828 0.13886 -18.064 < 2e-16 ***
---
Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2566.3 on 2029 degrees of freedom
Residual deviance: 2088.7 on 2027 degrees of freedom
AIC: 2094.7

Number of Fisher Scoring iterations: 4

> PseudoR2(red_mod1, which = c("CoxSnell", "Nagelkerke"))
  CoxSnell Nagelkerke
  0.2096727 0.2922123

> exp(coef(red_mod1)) # odds ratios
  (Intercept) X vocab$raceR1
  0.46684811 1.63274367 0.08140795

> confint(red_mod1) # confidence intervals
  2.5 % 97.5 %
  (Intercept) -1.1324260 -0.3936011
  X 0.4197295 0.5628576
  vocab$raceR1 -2.7854756 -2.2408354

> # reduced model (2)
> red_mod2 = glm(vocab$i3 ~ X, family=binomial(link="logit"))

> summary(red_mod2)
  Call:
  glm(formula = vocab$i3 ~ X, family = binomial(link = "logit"))

  Deviance Residuals:
    Min      1Q      Median      3Q      Max 
  -1.7270 -1.3446  0.7693  0.8886  1.2314 

  Coefficients:
  Estimate Std. Error z value Pr(>|z|)
  (Intercept) -0.29628 0.16471 -1.799 0.0721 .
  X 0.17030 0.02676 6.363 1.97e-10 ***
  ---
  Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

  (Dispersion parameter for binomial family taken to be 1)

  Null deviance: 2566.3 on 2029 degrees of freedom
  Residual deviance: 2525.1 on 2028 degrees of freedom
  AIC: 2529.1

  Number of Fisher Scoring iterations: 4

> PseudoR2(red_mod2, which = c("CoxSnell", "Nagelkerke"))
  CoxSnell Nagelkerke
  0.02010443 0.02801873

```

($-2\ln L_{f/R1} = 2088.7$) with reduced model (2) ($-2\ln L_{R2} = 2525.1$; X line) shows a significant ΔG^2 of 436.45. (As mentioned above, the comparison of the X predictor model with the constant only model (NULL line) will typically be significant in our context. Thus, we ignore the ΔG^2 of 41.23.) We use the PseudoR2 function to obtain our R^2 s for each of our models.

To obtain our parameter estimates, we use the summary function. The Coefficients tables for each model show that our models are

$$\begin{aligned}\text{full model: } \hat{z}_3 &= b_0 + b_1(X) + b_2 * \text{raceR1} + b_3(X * \text{raceR1}) \\ &= -1.17680 + 0.58303(X) - 1.69926 * \text{raceR1} - 0.16094 * X * \text{raceR1}\end{aligned}$$

$$\begin{aligned}\text{reduced model (1): } \hat{z}_3 &= b_0 + b_1(X) + b_2 * \text{raceR1} \\ &= -0.76175 + 0.49026(X) - 2.50828 * \text{raceR1}\end{aligned}$$

$$\begin{aligned}\text{reduced model (2): } \hat{z}_3 &= b_0 + b_1 X \\ &= -0.29628 + 0.1703(X)\end{aligned}$$

If we engaged in model selection, then given our results we reject the no DIF model (reduced model (2)). Moreover, because of the small effect size between the full model and the nonuniform DIF model ($\Delta R^2 = 0.0022$), we select the uniform DIF model (i.e., reduced model (1) / 'full' model) as our preferred model for item 3. Thus, this item exhibits uniform DIF and b_2 's sign indicates that Focal group members perform better than Reference group members. To obtain our corresponding odds ratio, we extract the regression coefficients (coef(red_mod1)) and exponentiate them (i.e., exp(. . .)). Thus, our odds ratio for raceR1 is OR = 0.081 (vocab\$raceR1 column).

Although several R packages implement the logistic regression approach, such as lordif (Choi, Gibbons, & Crane, 2011, 2016a, 2016b) and sirt's (Robitzsch, 2018) dif.logistic.function, we perform our analysis using difR's difLogReg function (Table 12.10). As above, we save our results to an output object (vocablr) in our call to difLogReg (vocablr = difLogReg(vocab, . . .)). The Logistic regression DIF statistic table contains $df = 2$ tests (Stat column) of the no DIF model (i.e., reduced model (2)) against the full model (i.e., $\Delta G^2 = (-2\ln L_{R2}) - (-2\ln L_F)$); see Endnote 8. As such, these reflect simultaneously testing for uniform and nonuniform DIF. All items are flagged as exhibiting significant DIF, with items 3 and 7, our two B items, exhibiting large effect sizes (we use JG effect size guidelines).¹⁰ As done above, given item 3's effect size, we remove it and repeat our analysis. After iteration 2, item 7 has both the largest effect size and a significant ΔG^2 , albeit all items have significant ΔG^2 s. Consequently, we remove item 7 and repeat our analysis. Iteration 3 shows that item 9 has both the largest effect size and largest ΔG^2 . Thus, we remove item 9 in our fourth iteration. Although some items still have significant ΔG^2 , our remaining seven items exhibit minimal DIF according to their corresponding effects sizes. As mentioned above, the removed items would be subjected to panel review to determine whether they should be considered to be biased.

TABLE 12.10. difR Analysis—Logistic Regression

```

> # This is a continuation of the session from Table 12.9

> library(difR)
> packageVersion("difR")
[1] '5.1'

> vocab=within(vocab,rm(raceR0))                                # remove ref=0 & focal=1 variable

> print((vocab$lr=difLogReg(vocab, group="raceR1", focal.name=0)))
  Detection of both types of Differential Item Functioning
  using Logistic regression method, without item purification
  and with LRT DIF statistic

  Matching variable: test score
  No set of anchor items was provided
  No p-value adjustment for multiple comparisons

  Logistic regression DIF statistic:
    Stat.      P-value
    i1      55.4362  0.0000 ***
    i2      76.8181  0.0000 ***
    i3     440.4723  0.0000 ***
    i4      22.2255  0.0000 ***
    i5      35.0503  0.0000 ***
    i6     16.6356  0.0002 ***
    i7     80.4288  0.0000 ***
    i8     27.6909  0.0000 ***
    i9    100.5045  0.0000 ***
   i10    43.2499  0.0000 ***

  Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Detection threshold: 5.9915 (significance level: 0.05)

  Items detected as DIF items:
  i1
  i2
  i3
  i4
  i5
  i6
  i7
  i8
  i9
  i10

  Effect size (Nagelkerke's R^2):

  Effect size code:
  'A': negligible effect
  'B': moderate effect
  'C': large effect

      R^2      ZT  JG
  i1  0.0213  A  A
  i2  0.0257  A  A
  i3  0.1911  B  C
  i4  0.0078  A  A
  i5  0.0305  A  A
  i6  0.0063  A  A

```

(continued)

TABLE 12.10. (*continued*)

```

i7 0.0738 A C
i8 0.0217 A A
i9 0.0414 A B
i10 0.0144 A A

Effect size codes:
Zumbo & Thomas (ZT): 0 'A' 0.13 'B' 0.26 'C' 1
Jodoin & Gierl (JG): 0 'A' 0.035 'B' 0.07 'C' 1

Output was not captured!

> vocablr$logitPar[3,] # item 3: constant & regr coefficients, Full model
  (Intercept)      SCORE      GROUP SCORE:GROUP
-2.8760537    0.4320908   1.6992581   0.1509402

> # iteration 2
> vocabpurfctn=within(vocab,rm(i3))
> print((vocablr=difLogReg(vocabpurfctn, group="raceR1", focal.name=0)))
:
Logistic regression DIF statistic:
  Stat.  P-value
i1  26.7854  0.0000 ***
i2  43.5120  0.0000 ***
i4  7.2335  0.0269 *
i5  19.5795  0.0001 ***
i6  29.0535  0.0000 ***
i7 108.6282  0.0000 ***
i8  11.0645  0.0040 **
i9 161.2790  0.0000 ***
i10 12.2610  0.0022 **

:
Effect size (Nagelkerke's R^2):
:
  R^2      ZT JG
i1 0.0101 A A
i2 0.0142 A A
i4 0.0024 A A
i5 0.0169 A A
i6 0.0109 A A
i7 0.1001 A C
i8 0.0086 A A
i9 0.0645 A B
i10 0.0040 A A
:

> # iteration 3
> vocabpurfctn=within(vocabpurfctn,rm(i7))

> print((vocablr=difLogReg(vocabpurfctn, group="raceR1", focal.name=0)))
:
Logistic regression DIF statistic:
  Stat.  P-value
i1  20.8913  0.0000 ***
i2  37.3749  0.0000 ***
i4  7.0063  0.0301 *
i5  16.5115  0.0003 ***
i6  30.5332  0.0000 ***
i8  8.4544  0.0146 *
i9 185.6361  0.0000 ***
i10 6.9135  0.0315 *

```

(continued)

TABLE 12.10. (*continued*)

```

:
Effect size (Nagelkerke's R^2):
:
R^2      ZT JG
i1  0.0078 A  A
i2  0.0120 A  A
i4  0.0023 A  A
i5  0.0142 A  A
i6  0.0114 A  A
i8  0.0066 A  A
i9  0.0730 A  C
i10 0.0022 A  A
:

> # iteration 4
> vocabpurfctn=within(vocabpurfctn,rm(i9))
> print((vocabblr=difLogReg(vocabpurfctn, group="raceR1", focal.name=0)))
:
Logistic regression DIF statistic:
  Stat.   P-value
i1 16.0512  0.0003 ***
i2 37.8604  0.0000 ***
i4 12.9530  0.0015 **
i5  9.4449  0.0089 **
i6 41.8897  0.0000 ***
i8  4.9609  0.0837 .
i10 4.4408  0.1086

:
Effect size (Nagelkerke's R^2):
:
R^2      ZT JG
i1  0.0060 A  A
i2  0.0120 A  A
i4  0.0043 A  A
i5  0.0080 A  A
i6  0.0153 A  A
i8  0.0038 A  A
i10 0.0014 A  A
:
```

Summary

Differential item functioning occurs when performance on an item is a function not only of person location, but also of tangential factor(s). DIF may be conceptualized as occurring when an item's response function changes across different groups of respondents. The groups are typically referred to as the Reference and Focal groups. The Focal group is the one being investigated to see if it is disadvantaged (or advantaged) by the item and, in general, is the "minority" group. The Reference group is the comparison group and, in general, is the "majority" group. DIF methods not only detect the presence of DIF, but also whether the Focal or Reference group is favored.

There are two forms of DIF. In one form, uniform DIF, one group performs better

than the other group throughout the continuum. Graphically, uniform DIF is represented as a Reference group IRF that is parallel to the Focal group IRF. In the other form, nonuniform DIF, the Reference group performs better than the Focal group for a particular portion of the continuum, whereas along a different portion of the continuum the Focal group performs better than the Reference group. Graphically, nonuniform DIF is represented as a Reference group IRF that crosses the Focal group IRF. In short, for nonuniform DIF there is an interaction between performance on the item, group membership, and location along the latent continuum, whereas for uniform DIF there is no interaction.

Several approaches can be used to identify whether an item is exhibiting DIF. Some of these approaches are IRT-based (e.g., TSW- ΔG^2), whereas others are not (e.g., Mantel-Haenszel Chi-Square, logistic regression).

If an item is identified as exhibiting meaningful DIF, this does not automatically mean that the item is biased. For an item to be considered biased, the DIF item is subjected to review by a panel of experts to determine whether the source of an item's differential performance is relevant or irrelevant to the construct being measured by the instrument. It is the panel's conclusion that determines whether an item exhibiting DIF is also biased. As a consequence, if a researcher/practitioner is developing an instrument, then they may simply be concerned with whether one or more items are exhibiting significant DIF and not whether any DIF items are also biased. In this case there would not be a need to establish a formal bias review panel.

The researcher/practitioner should take into account the magnitude of the DIF (i.e., effect size), statistical significance vis-à-vis the sample size, as well as the number of different Reference/Focal group comparisons that were used in the DIF analysis (i.e., male vs. female, Racial Group 1 vs. Racial Group 2, Ethnic Group 1 vs. Ethnic Group 2, etc.) before making decisions about removing items exhibiting DIF. For instance, because statistical tests are influenced, in part, by the sample size, it is possible to obtain a statistically significant DIF statistic with a very large sample size even though the magnitude of the DIF is negligible. To avoid disadvantaging a manifest group, DIF analyses should use all manifest groupings that are relevant for the population to which the instrument will be administered.

One potential advantage of the logistic regression approach over the TSW- ΔG^2 approach is that logistic regression does not require any of the assumptions underlying IRT. With respect to the MH technique, the logistic regression approach allows an examination of both uniform and nonuniform DIF. Whether the $MH\chi^2$ would flag an item for nonuniform DIF depends on the amount of cancellation that takes place. In short, the $MH\chi^2$ should only be used for detecting uniform DIF. An additional potential advantage of the logistic regression DIF method over the MH approach is that unlike the MH statistic's use of discrete summed scores for conditioning, logistic regression also allows the use of a nondiscrete person location predictor. Unlike the other approaches mentioned, the logistic regression procedure allows for the use of covariates. In Chapter 13 we use a multilevel approach to investigate DIF.

Notes

1. This chapter is concerned with changes in item parameter estimates *across groups*. However, a different form of DIF involves changes in item parameter estimates *over time*. This type of DIF is known as *item parameter drift* or *item drift* (Thissen et al., 1988). To investigate item drift, one may use the Thissen et al. (1988) likelihood ratio ($TSW-\Delta G^2$) approach discussed in this chapter.
2. The program IRTDIF (Kim & Cohen, 1992) and the R packages `difR` (Magis, 2018; Magis, Béland, Tuerlinckz, & De Boeck, 2010) and `DFIT` (Cervantes, 2017a, 2017b) can be used to obtain Lord's Chi-Square, the Exact Signed Area, and H Statistic approaches.
3. In the context of the $TSW-\Delta G^2$, the procedure's step 1 yields the full model (i.e., $G_1^2 = G_F^2$), whereas step 2 results in the reduced model (i.e., $G_2^2 = G_R^2$). Therefore,

$$\Delta G^2 = G_2^2 - G_1^2 = G_R^2 - G_F^2$$

The log likelihood statistics are

$$G_2^2 = G_R^2 = -2 \ln(L_R)$$

and

$$G_1^2 = G_F^2 = -2 \ln(L_F).$$

Then, by substitution, the likelihood ratio test is

$$\Delta G^2 = G_R^2 - G_F^2 = (-2 \ln L_R) - (2 \ln L_F))$$

4. Although DIF analyses occur in a nonexperimental setting, we refer to the product of group membership and a person's location as their *interaction* rather than as their *joint relationship* (see Pedhauzer, 1997).
5. Because this is a general procedure, multiple statistics are available to us. For example, when we have more than two rows or columns and the tables are not inherently ordered, then the Nonzero Correlation and General Association values will differ. In this case, General Association value is used.
6. Our coding reflects using 0 to represent a reference group, as is done in some statistical techniques such as multiple regression. However, coding the Reference group as 1 and the Focal group as 0 eliminates the need to take the reciprocal of the output's common odds ratio estimate.
7. In SPSS our syntax would be

```
CROSSTABS  
/TABLES=race BY i3 BY X  
/FORMAT=AVALUE TABLES  
/STATISTICS=CMH(1)  
/CELLS=COUNT  
/COUNT ROUND CELL.
```

The corresponding output follows. As can be seen, SPSS uses the continuity correction.

		Crosstab		Total	
		Count			
		X			
1.00	race	0	0	1	
2.00	race	1	6	8	
		Total	7	9	
10.00	race	0	7	7	
		1	24	42	
		Total	31	49	
		:			
10.00	race	1	0	35	
		Total	16	130	
Total	race	0		13	
		1		3	
		Total		16	
Total	race	0	503	1022	
		1	161	1008	
		Total	664	1366	
				2030	

Tests of Homogeneity of the Odds Ratio

Chi-Squared	df	Asymptotic Significance (2-sided)
Breslow-Day	8.948	8 0.347
Tarone's	8.886	8 0.352

Tests of Conditional Independence

Chi-Squared	df	Asymptotic Significance (2-sided)
Cochran's	384.579	1 0.000
Mantel-Haenszel	381.089	1 0.000

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate	12.205
ln(Estimate)	2.502
Standard Error of ln(Estimate)	0.144
Asymptotic Significance (2-sided)	0.000

Asymptotic 95% Confidence Interval

Common Odds Ratio	Lower Bound	9.204
Upper Bound		16.186
ln(Common Odds Ratio)	Lower Bound	2.220
Upper Bound		2.784

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Under the conditional independence assumption, Cochran's statis-

tic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate
Estimate 12.205
ln(Estimate) 2.502
Standard Error of ln(Estimate) 0.144
Asymptotic Significance (2-sided) 0.000

Asymptotic 95% Confidence Interval
Common Odds Ratio Lower Bound 9.204
Upper Bound 16.186
ln(Common Odds Ratio) Lower Bound 2.220
Upper Bound 2.784

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

The Estimate and ln(Estimate) values in the Mantel-Haenszel Common Odds Ratio Estimate table can be used to obtain α_{MH} and β_{MH} . Specifically, given our coding and how it corresponds to how SPSS calculates its common odds, we take the reciprocal of the Estimate value ($\hat{\alpha}_{MH} = 1/\text{Estimate} = 1/12.205 = 0.0819$) and $\hat{\beta}_{MH} = -l^*\ln(\text{Estimate}) = -2.5019$.

In SYSTAT our command is

```
> PLENGTH NONE / FREQ MANTEL  
> TABULATE X * RACE * I3
```

The corresponding output follows. As can be seen, SYSTAT also uses the continuity correction.

```
Crosstabulation: Multiway: Tabulate  
Counts*  
X = 1  
    RACE(rows) by I3(columns)  
        0      1      Total  
    0      1      0      1  
    1      6      2      8  
    Total  7      2      9  
  
X = 2  
    RACE(rows) by I3(columns)  
        0      1      Total  
    0      7      0      7  
    1     24     18     42  
    Total 31     18     49  
:
```

```

X = 10
RACE(rows) by I3(columns)
      0      1    Total
0     0     13     13
1     0      3      3
Total 0     16     16

Mantel-Haenszel Statistic : 12.2054
Mantel-Haenszel Chi-Square : 381.08889
p-Value                   : 0

```

Analogous to what is done above, the Mantel-Haenszel Statistic can be used to obtain $\hat{\alpha}_{MH}$ and $\hat{\beta}_{MH}$. That is, $\hat{\alpha}_{MH} = 1/12.205 = 0.0819$ and $\hat{\beta}_{MH} = \ln(\alpha_{MH}) = -2.5019$.

8. We could compare the full model's $-2\ln L$ with that of reduced model (2) to simultaneously test for uniform and nonuniform DIF. This would be a two *df* test, with the significance level reduced to, say 0.01, to account for the multiple hypotheses being tested (Zumbo, 1999). For our example, this test would be significant with a value of

$$\Delta G^2 = (-2\ln L_{R2}) - (-2\ln L_F) = 2525.103 - 2084.631 = 440.472$$

Furthermore, $\Delta R^2 = R_F^2 - R_{R2}^2 = 0.2944 - 0.0280 = 0.2664$. Therefore, we have evidence of significant DIF associated with a large effect size.

9. As an example of the interpretation of the coefficient estimates in the model, assume the reduced model (1)/'full' model is the model of interest (i.e., the interaction term in the full model did *not* lead to a significant improvement in fit). From Table 12.6 our model is

$$\text{reduced model (1): } \hat{z}_3 = 0 - .7618 + 0.4903(X) - 2.5083 * \text{raceR1}$$

Therefore, holding the summed score fixed and switching from the Focal group to the Reference group result in a *decrease* in the log odds of obtaining a response of 1 by 2.5083 after controlling for X . In terms of odds we have that the odds that a Reference group member will produce a response of 1 are $\exp(-2.5083) = 0.0814$ to 1 (note: 0.081 is the value listed as the Point Estimate for raceR1 in the Odds Ratio Estimates table). Alternatively, holding the summed score fixed, one expects the odds of Focal group members to correctly respond to the item to be roughly 12 to 1 (i.e., $1/0.0814 = 12.284$) relative to comparable Reference group members.

10. Although difR labels its effect size table as Effect size (Nagelkerke's R^2). These values correspond to using the Cox and Snell (1989) R^2 and not Nagelkerke R^2 in calculating ΔR^2 . For example, from Table 12.10 and item 3, we have for the comparison of the full model with the reduced model (2) $\Delta R^2 = 0.1911$. Using the values from Tables 12.6 and 12.7 for the Nagelkerke R^2 s (i.e., Max-rescaled R-Square), we obtain

$$\Delta R^2 = R_F^2 - R_{R2}^2 = 0.2944 - 0.0280 = 0.2664.$$

(See Example: DIF Analysis of vocabulary test, SAS proc logistic.) However, if we use the Cox and Snell R^2 (R-Square), we have

$$\Delta R^2 = R_F^2 - R_{R2}^2 = 0.2112 - 0.0201 = 0.1911$$

For completeness the Cox and Snell R^2 is

$$R_{\text{Cox\&Snell}}^2 = 1 - \left(\frac{L_{b0}}{L_{\text{Model}}} \right)^{\frac{2}{N}}, \quad (12.9)$$

where L_{b0} is the likelihood for the null (i.e., the constant only) model and L_{Model} is the likelihood for the model of interest. Because the Cox and Snell R^2 has a maximum value of $1 - (L_{b0})^{2/N}$, Nagelkerke (1991) rescaled it so that R^2 could achieve an upper limit of 1.

$$R_{\text{Nagelkerke}}^2 = \frac{R_{\text{Cox\&Snell}}^2}{1 - (L_{b0})^{\frac{2}{N}}}. \quad (12.10)$$

13

Multilevel IRT Models

In this chapter, we discuss an alternative way of conceptualizing our IRT models. In the following, we conceptualize our IRT models from this multilevel perspective; a multilevel conceptualization of CTT may be found in Miyazaki (2005) and Miyazaki and Skaggs (2008). This multilevel conceptualization extends our models to include factors (e.g., schools, countries, organizations) that can account for variability in our items and/or persons.¹ Multilevel IRT models have been used to identify item dependence, detect differential item and testlet functioning, incorporate response time, model multidimensionality, and handle nonresponses (e.g., Adams, Wilson, & Wang, 1997; Beretvas & Walker, 2012; Cho, Gilbert, & Goodwin, 2013; Jiao, Wang, & Kamata, 2005; Klein Entink, Kuhn, Hornke, & Fox, 2009; Maier, 2002; Pastor & Beretvas, 2006; Randall, Cheong, & Engelhard, 2011).

Multilevel IRT—Two Levels

In Chapter 4, we used an ANOVA framework in our discussion of marginal maximum likelihood. Specifically, we considered the repeated measures (within-subjects) design in which we had repeated observations on each of our randomly sampled participants. In this design, a participant's multiple observations are nested within the participant. In a similar fashion, a respondent's responses to a series of items reflect multiple observations of the respondent. Consequently, item responses are nested within the respondent. Stated another way, respondents and their responses reflect different levels in our data's structure. In the simplest case, we have two levels in which responses comprise level 1 and respondents represent level 2. With multilevel IRT our level 2 can also include item-oriented and/or person-oriented predictors. (Below we present examples of these predictors.) Additionally, we can extend this idea to have respondents nested within one or more background variables, such as countries, schools, and classrooms, to produce a three-level situation (e.g., items within persons within countries). For instance, we may be using IRT in a longitudinal context. In this case, levels 1 and 2 would represent items

and people, respectively, with level 3 reflecting the time factor (e.g., see Huang, 2015; Pastor & Beretvas, 2006). We begin with a two-level model.

Multilevel models consist of a series of regression models in which lower-level model parameter(s) serve as the criterion variable(s) for higher-level models. For the lowest level model, the item response is the criterion. As such, we have a hierarchical set of linear regression models or hierarchical linear models (HLMs). However, because this latter term is easily confused with hierarchical regression models (i.e., a successive set of nested models such as Chapter 12's full, reduced model (1) and reduced model (2)) we use the term *multilevel*.²

In the simplest case, our item score will be dichotomous. The sampling (conditional) distribution of our dichotomous item response, given some probability of a response of 1, cannot reasonably be assumed to be normally distributed. Rather, a single-item response follows a binomial distribution given some probability of a response of 1; the item response is a random variable. This one trial binomial distribution is a Bernoulli distribution. The probability of a response of 1, p_{ij} , is given by one of our dichotomous models such as the 1PL model. In the following, we assume the Rasch/1PL model, but the ideas can be extended to non-Rasch models.³

Because the multilevel model represents a series of regression models, the regression notational system (or some variant thereof) is typically used. Therefore, we transition from our α , δ_j , and γ_j notation to a regression notation. In Chapter 2, we first presented our 1PL model in a slope–intercept format. This slope–intercept parameterization of the model's exponent for item j is $\vartheta_{ij} = \alpha(\theta_i - \delta_j) = \alpha\theta_i + \gamma_j$ where the intercept (constant) $\gamma_j = -\alpha\delta_j$. In terms of regression's beta notation, we have

$$\vartheta_{ij} = \gamma_j + \alpha\theta_i = \beta_{0j} + \beta_{1j}\theta_i, \quad (13.1)$$

where β_{0j} is our intercept (constant, γ_j) and β_{1j} is our slope (regression coefficient, α) for item j . In the case of the Rasch model in which $\alpha = \beta_{1j} = 1$, then $\beta_{0j} = \gamma_j = -\delta_j$ and we have the log odds of a response of 1 is

$$\ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = \vartheta_{ij} = \beta_{0j} + \beta_{1j}\theta_i = \beta_{0j} + \theta_i. \quad (13.2)$$

As in previous chapters, we assume item responses are independent conditional on a person's location. However, in the multilevel context this is rephrased as conditional independence means item responses (level 1) are assumed independent of one another when nested within respondent (level 2). Implied in our use of the Rasch model is that we are using the logit link function; see Appendix G, “Odds, Odds Ratios, and Logits.”⁴ In Equation 13.2, our slope and intercept parameters are being considered to be fixed across respondents (i.e., items are fixed effects).

Equation 13.2 consists of a person component and an item component. Extending Equation 13.2 to more than a single item, our regression model contains a categorical item predictor variable in addition to our person component. The standard regression approach to capture the information from a categorical predictor is to create as

many indicator “variables” as there are *dfs* in the categorical predictor.⁵ Therefore, for a *L*-item instrument and following Kamata (2001; also see Adams, Wilson, & Wu, 1997; De Boeck & Wilson, 2004), we have for person *i* and item *j* the regression model

$$\vartheta_{ij} = \beta_{0i} + \beta_{1i}X_{1ij} + \beta_{2i}X_{2ij} + \dots + \beta_{(L-1)i}X_{(L-1)ij}, \quad (13.3)$$

where, $X_{1ij}, X_{2ij}, \dots, X_{(L-1)ij}$, are indicator (predictor) “variables.” Letting $X_{(q)ij}$ be the *qth* indicator variable for $q = 1 \dots (L - 1)$, then when $q = j$ we have $X_{(q)ij} = 1$ and when $q \neq j$ we have $X_{(q)ij} = 0$. Thus, the $X_{(q)ij}$ s simply indicate the administration of an item. Unpacking Equation 13.3 for an *L*-length instrument for person *i*, we have for items 1, 2, and so on

$$\vartheta_{i1} = \beta_{0i} + \beta_{1i}. \quad (13.4)$$

$$\vartheta_{i2} = \beta_{0i} + \beta_{2i}. \quad (13.5)$$

⋮

$$\vartheta_{i(L-1)} = \beta_{0i} + \beta_{(L-1)i}. \quad (13.6)$$

We can collectively represent Equations 13.4 to 13.6 for person *i* as

$$\ln \left[\frac{p_{ij}}{1-p_{ij}} \right] = \vartheta_{ij} = \beta_{0i} + \sum_{q=1}^{L-1} \beta_{qi} X_{qij}, \quad (13.7)$$

where the (partial) regression coefficient β_{qi} is the *qth* item effect (i.e., the unique effect of the *qth* indicator vector) with respect to the constant β_{0i} .^{6,7} In this context, β_{qi} reflects an offset or difference from the constant β_{0i} for person *i*. We can interpret our constant β_{0i} as the overall mean effect across items or, alternatively, the effect that is common to all items for person *i*. As such, it reflects a baseline for person *i*. The upper summation limit in Equation 13.7 reflects the item variable’s *df*. Thus, the *Lth* item’s expected effect for person *i* is given by β_{0i} ’s value because we are using the *Lth* item as the reference (baseline) item to identify the model. In general, for person *i* and item *j* (i.e., $q = j$), the log odds of a response of 1 is

$$\ln \left[\frac{p_{ij}}{1-p_{ij}} \right] = \vartheta_{ij} = \beta_{0i} + \beta_{qi}. \quad (13.8)$$

Equation 13.8 represents our *item-level* model (level-1) in which our items are nested within person *i*. Figure 13.1 contains a conceptual representation of Equation 13.8 for two respondents and items 1, 2, through the *L-1* item; at present we focus only on Persons 1 and 2. As can be seen, our person baseline locations β_{0i} s are not (necessarily) located at the same point. Moreover, our item offsets β_{qi} s can vary from one another but are constant across persons. That is, person 1’s offset, β_{11} , equals person 2’s offset, β_{12} ; this is also true for the remaining items $\beta_{21} = \beta_{11}, \dots, \beta_{(L-1)1} = \beta_{(L-1)2}$.

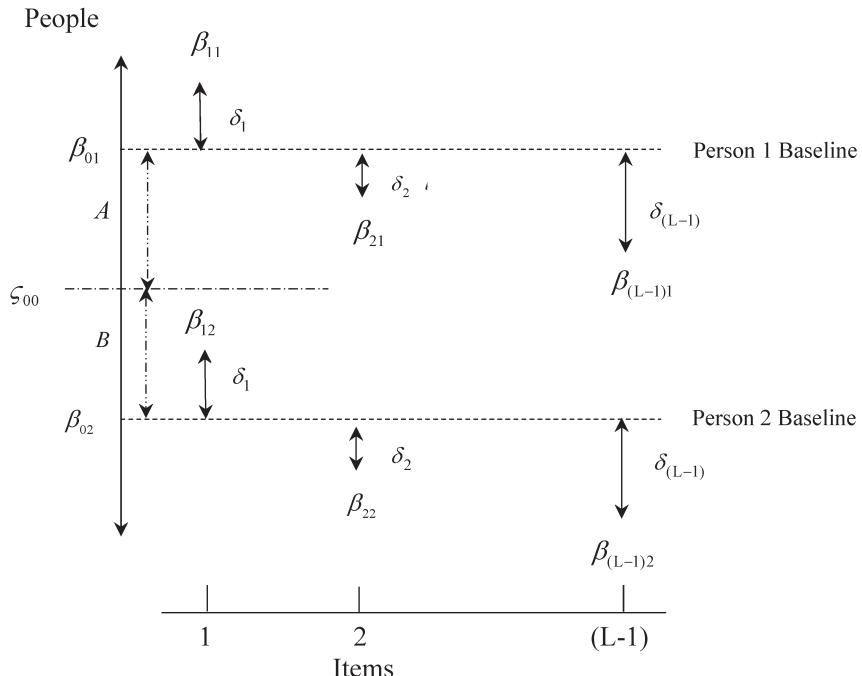


FIGURE 13.1. Conceptual representation of modeling.

We can view Equation 13.8 as modeling person i 's “performance” on item j because it contains the person intercept parameter, β_{0i} , that reflects an overall effect shared across items (i.e., what is shared across items is related to θ_i ; see Equation 13.11) and an item effect parameter, β_{qi} , that reflects the difference from β_{0i} on item j . Therefore, the probability that person i provides a response of 1 on item j is

$$P_{ij} = \frac{e^{\vartheta_j}}{1+e^{\vartheta_j}} = \frac{e^{\beta_{0i} + \beta_{qi}}}{1+e^{\beta_{0i} + \beta_{qi}}} \quad (13.9)$$

At level 2, we have our *person-level* model in which we predict our item-level model's constant/intercept and regression coefficient parameters, β_{0i} and β_{qi} , respectively. Recall that with respect to the β_{qi} 's each item's effect is considered to be constant across respondents (i.e., item is a fixed effect). That is, for all respondents, item 1 is located at the same location (e.g., 1.5); for all respondents item 2 is located at the same location (e.g., -2); and so on. As such, because for an item there is no item variation to model, there is no random effect associated with the items in our person-level models.⁸ This characteristic is reflected in Figure 13.1 by $\beta_{11} = \beta_{12}$, $\beta_{21} = \beta_{11}, \dots, \beta_{(L-1)1} = \beta_{(L-1)2}$. Therefore, at level 2 we have for person i and our $L - 1$ item regression coefficients

$$\beta_{1i} = \varsigma_{10}$$

$$\beta_{2i} = \varsigma_{20}$$

⋮

$$\beta_{(L-1)i} = \zeta_{(L-1)0}, \quad (13.10)$$

or, in general, $\beta_{qi} = \zeta_{q0}$ where ζ 's second subscript, 0, indicates that it is a level-2 constant. (See De Boeck (2008) for why and when items may be considered random.)

As mentioned above, the constant in our model for person i , β_{0i} , represents the effect of the L th item for person i or, stated another way, the overall mean effect across (or common to) items for person i . Because different respondents (e.g., high- and low-ability respondents) may have different mean effects, we can model this variability. (This variability is shown in Figure 13.1 by the difference between β_{01} and β_{02} (i.e., $A + B$.) We can try to account for these distances by using a model that contains only the mean of the β_{0i} 's ζ_{00} (in Figure 13.1) because, all things being equal, our best prediction is the mean. The variability remaining after using the mean is captured by the residual term θ_{0i}^r . This remaining variability reflects differences among the respondents that is due to only their differing locations on the latent variable. (With complicated models, the differences may also be due to additional factor(s).) Symbolically, person i 's level-1 baseline is modeled at level 2 by fitting the constant, ζ_{00} , with the residual variance reflecting the random person effect (θ_{0i}^r)

$$\beta_{0i} = \zeta_{00} + \theta_{0i}^r, \quad (13.11)$$

where the residual term, θ_{0i}^r , is assumed to be distributed as $N(0, \sigma_\theta^2)$. Thus, person i 's average overall effect across items is a function of the average performance of all respondents (ζ_{00} , the average overall effect across all persons and items) plus person i 's specific characteristic(s) (e.g., their location on the latent variable). Because we are modeling the variability in β_{0i} respondents are a random effect (i.e., β_{0i} is allowed to vary across individuals). We have an Equation 13.11 for each of our N respondents across which our θ_{0i}^r 's vary (i.e., $\theta_{01}^r, \theta_{02}^r, \dots, \theta_{0N}^r$ can vary).

Our item- and person-level models can be combined into a single model by substituting Equations 13.10 and 13.11 into Equation 13.1. Therefore, for person i and item j (i.e., $j = q$) the log odds of a response of 1, we have

$$\vartheta_{ji} = \beta_{0i} + \beta_{qi} = \zeta_{00} + \theta_{0i}^r + \zeta_{q0}. \quad (13.12)$$

Equation 13.12 states that person i 's log odds of success on item j is due to a common effect to all items (ζ_{00}), item j 's specific effect (ζ_{q0} when $q = j$), and the person's latent characteristic of interest (θ_{0i}^r ; e.g., proficiency). Our item's location (in terms of easiness) is

$$\delta_j^E = (\zeta_{j0} + \zeta_{00}). \quad (13.13)$$

We can rearrange Equation 13.12 into a difference format to conform to the typical IRT representation. To do so, we first convert our easiness item location to be an item location in terms of difficulty

$$\delta_j = -1^*(\delta_j^E) = -1^*(\zeta_{j0} + \zeta_{00}) = (-\zeta_{j0} - \zeta_{00}). \quad (13.14)$$

Therefore, Equation 13.12 becomes

$$\vartheta_{ji} = \zeta_{00} + \theta_{0i}^r + \zeta_{q0} = \theta_{0i}^r - (-\zeta_{q0} - \zeta_{00}). \quad (13.15)$$

By substitution we obtain the probability of a response of 1 by person i on item j

$$p_{ji} = \frac{\exp[\vartheta_{ji}]}{1 + \exp[\vartheta_{ji}]} = \frac{\exp[\theta_{0i}^r - (-\zeta_{j0} - \zeta_{00})]}{1 + \exp[\theta_{0i}^r - (-\zeta_{j0} - \zeta_{00})]} = \frac{\exp[\theta_{0i}^r - \delta_j]}{1 + \exp[\theta_{0i}^r - \delta_j]}. \quad (13.16)$$

Equation 13.16 is our 1PL model with $\alpha = 1$ (i.e., Rasch model⁹) and shows its conceptualization as a two-level model in which one level represents items “nested” within a second (person) level. As is true with the 1PL/Rasch model, all respondents with the same observed score obtain the same location estimate. Our θ_{0i}^r differs from θ_i in that the former is considered a normally distributed random variable with mean 0 and a variance of $\sigma_{\theta_i}^2$, whereas the latter is considered to be a fixed variable (e.g., when using JMLE/CMLE) or to be random (e.g., when using MMLE θ_i). (The normality assumption for θ_{0i}^r is in addition to those presented in Chapter 2. Moreover, when our persons are nested within another factor (e.g., schools, programs, countries), it is also assumed that our latent variable is normally distributed between this factor’s levels.)

Of course, the observed variability in the β_{0i} depicted in Figure 13.1 may not be due solely to person locations. We can account for whatever variability remains after fitting the mean by using respondent characteristics such as a respondent’s opportunity to learn or their gender, and so on. These additional characteristics would be level-2 model predictors.

Through Equations 13.10 and 13.11, we may also include additional predictors that can be combined into a single model reflecting both person and item levels. For instance, we may choose to predict our item effects, β_{qi} s, by including item characteristics, as is done in the linear logistic test model (Fischer, 1973). Alternatively (or in addition to including item characteristics), we may add person-oriented predictor variable(s), such as anxiety level, background, response time, gender, race, emotionality, and so on, to our model (e.g., Casabianca, Junker, Nieto, & Bond, 2017; Fox, 2004; Muckle & Karabatsos, 2009). After demonstrating the equivalence of the Rasch model and its multilevel parameterization, we discuss the use of person-oriented predictors for items. For instance, does a person-oriented predictor, such as gender, affect our item location estimates (i.e., DIF)? Additionally, we discuss person-oriented predictors for respondents. For example, does educational level affect a respondent’s nutritional literacy?

Example: Estimating the Rasch Model from a Multilevel Perspective, proc glimmix

In Chapter 2, we introduced the response data from a five-item mathematics examination administered to over 19,000 examinees. For this example, we sample 1000 examinees.

ees from our original data set. Our sample has a mean observed score of 2.53 ($SD = 1.08$) and minimum and maximum observed scores of 1 and 4, respectively.

Various specialized packages and programs are available for performing multi-level analyses, for example, HLM (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011), MLwiN (and R2MLwiN) (Charlton, Rasbash, Browne, Healy, & Cameron, 2017), Mplus, and the R packages `lme4` (Bates, Mächler, Bolker, & Walker, 2015), `mirt` (i.e., `mixedmirt`), `multilevel` (Bliese, 2016), and `nlme` (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018). (A comparison of five multilevel modeling packages was performed by McCoach et al. (2018).) We will first use SAS proc `glimmix` for our analysis and then `lme4`.¹⁰

For previous clibrations, our response data were in an N by L format, or what some would refer to as an “unstacked” (or “wide”) format. However, to perform our multi-level analyses, we need to have all the row-oriented response vectors transposed and “stacked” into a single column-oriented response variable; this variable is in row-(major) order. Therefore, each response vector consists of multiple rows, with the response vectors concatenated across people. Organized this way, the data are in a “stacked” format (a.k.a., “long” format).

Figure 13.2 shows a schematic of the transformation of unstacked data to the corresponding stacked data. As can be seen, we have a `person` variable containing an examinee ID to which we add an `item` variable, our response variable `x`, as well as our design matrix (X_{1i1}, \dots, X_{5i5}). Our unstacked data contain 1000 rows (each row representing an examinee’s responses to the five items), whereas the stacked data set contains 5000 rows (1000 persons x 5 items each). The first five rows of the stacked format contain the first examinee’s responses in `x`, with the `person` variable listing the corresponding ID value of 1 and the `item` variable’s value representing the item to which the response corresponds. We also present the indicators (X_{1i1}, \dots, X_{5i5}) to show the corresponding coding; the fifth indicator (the L th item) is identified by a code of 0 on each of the previous four indicators. As such, it is only necessary to provide $L-1$ indicator predictors. The second set of five rows contains the analogous information for the second examinee and so on.

Table 13.1 contains the data steps and `glimmix` procedure commands. In SAS, converting unstacked data to a stacked format requires two data steps and two data sets (i.e., `unstacked_data`, `stacked_data`). The first data step reads our math data set and stores it in the `unstacked_data` set, whereas our second data step performs the conversion from unstacked to stacked format. In the latter step, we use an array to collectively treat our five items. Thus, we define our `item_array` to be an array of length $L = 5$ and assign it our five item variables (`array item_array(1:5) i1-i5`). In our `do` statement we traverse the length of our array one element at a time, assigning each element’s value to a variable `x`, and after each assignment, we save (output) the corresponding value to the second data set (`stacked_data`). At the end of our second data step, we delete the now superfluous item variables `i1-i5` (`drop i1-i5`).¹¹

In our `proc` step, we specify the use of the data set `stacked_data` (`data = stacked_data`) and the (adaptive) quadrature estimation (`method =`

Original (unstacked) Format		Stacked Format ^a						
Person		Person	item	x				
1	1 1 0 0 0	1	1	1	0	0	0	0
2	1 1 1 0 0	1	2	1	0	1	0	0
:		1	3	0	0	0	1	0
1,000	1 1 1 1 0	1	4	0	0	0	0	1
		1	5	0	0	0	0	1
		2	1	1	1	0	0	0
		2	2	1	0	1	0	0
		2	3	1	0	0	1	0
		2	4	0	0	0	0	1
		2	5	0	0	0	0	1
		:						
		1000	1	1	1	0	0	0
		1000	2	1	0	1	0	0
		1000	3	1	0	0	1	0
		1000	4	1	0	0	0	1
		1000	5	0	0	0	0	1

FIGURE 13.2. Schematic of a data input file with five items.

^aThe item variable is not necessary if one provides the indicator predictors (i.e., the design matrix). However, its inclusion facilitates legibility by identifying the rows that are nested within a person. Its use with SAS class statement eliminates the need to provide the design matrix in the data set (i.e., all one needs is Person, item, and x). With other programs the item variable can be used to create the design matrix through variable creation and recoding. We provide the design matrix to show its layout for use with estimation programs (e.g., HLM) that require a design matrix. Although the design matrix requires only L – 1 columns for pedagogical reason we show it with L columns.

quadrature) on the proc statement; the bottom panel shows the SAS program using the indicator predictors. Use of the class statement instructs SAS to create the corresponding dummy variables for items (class item;). Our model statement specifies our criterion variable, x, to model the response code of 1 using the event syntax (event = '1') and the use of the logit link function, and affirms that we are dealing with a binary response distribution. In addition, this is where we specify our fixed effects predictor (item). We use the random statement to specify the effect that is allowed to vary; person is our random effect. To obtain our item easiness estimates, we include the lsmeans statement with the ilink option for items (i.e., we use the inverse link function to transform the logit estimates to a probability scale and print them and their standard errors).

Table 13.2 shows the output for our analysis. As can be seen, the Model Information table shows that we are using the logit link function, our estimation approach, and that our response distribution is binary. Use of the class statement results

TABLE 13.1. proc glimmix for Rasch Calibration

SAS data steps for reading unstacked data and converting it to a stacked format

```
data unstacked_data; /* data step #1 */
  infile "C: mathunstacked.dat";
  input i1-i5;
  person=_N_;
run;

data stacked_data; /* data step #2 */
  set unstacked_data;
  array item_array(1:5) i1-i5;
  do item=1 to 5;
    x=item_array(item);
    output;
  end; /* end do */

  drop i1-i5;
run;
```

proc step: Using the CLASS statement to generate the indicator variables and the response syntax in lieu of the events/trials syntax.^{a,b,c,d}

```
proc glimmix data=stacked_data method=quadrature;
  title "MM-Rasch formulation demo";
  class item;
  model x(event='1')= item / dist=binary link=logit;
  random intercept / subject=person type=un G solution;
  lsmeans item / ilink;
run;
```

Alternative proc step: Using the design matrix and use of the descending option in lieu of the event option.^b

```
proc glimmix data=stacked_data method=quad;
  title "MM-Rasch formulation demo, design matrix";
  model x(descending)= x1 x2 x3 x4 / dist=binary link=logit covb s;
  random intercept / subject=person type=un G solution;
run;
```

^aInclusion of the solution option on the model line produces the fixed effects estimates.

^bWhen using the binary distribution (dist=binary) SAS will by default model the 0 category as the event of interest. Therefore, with either the events/trials syntax or response syntax it is necessary to specify modeling the 1 category (i.e., event='1' or descending).

^cproc glimmix can also be used with polytomous ordered responses.

^dmethod=quadrature can be time and memory intensive. SAS/STAT 14.1 and newer versions support the use of the FASTQUAD option (i.e., method= quadrature(FASTQUAD qpoints=4). Alternatively, we could use method=laplace.

in the Class Level Information table that we should check to ensure the correct number of items is being used; alternatively, this information can be obtained from the Dimensions table; the Dimensions table presents the size of relevant matrices. The Number of Observation Read reflects the number of lines read from the stacked data file (N * L). After the Iteration History table, we see that we obtained con-

TABLE 13.2. proc glimmix Rasch Output

CLASS statement & random P																																																																									
The GLIMMIX Procedure																																																																									
Model Information																																																																									
<table border="1"> <tr><td>Data Set</td><td>WORK.D1</td></tr> <tr><td>Response Variable</td><td>x</td></tr> <tr><td>Response Distribution</td><td>Binary</td></tr> <tr><td>Link Function</td><td>Logit</td></tr> <tr><td>Variance Function</td><td>Default</td></tr> <tr><td>Variance Matrix Blocked By</td><td>person</td></tr> <tr><td>Estimation Technique</td><td>Maximum Likelihood</td></tr> <tr><td>Likelihood Approximation</td><td>Gauss-Hermite Quadrature</td></tr> <tr><td>Degrees of Freedom Method</td><td>Containment</td></tr> </table>		Data Set	WORK.D1	Response Variable	x	Response Distribution	Binary	Link Function	Logit	Variance Function	Default	Variance Matrix Blocked By	person	Estimation Technique	Maximum Likelihood	Likelihood Approximation	Gauss-Hermite Quadrature	Degrees of Freedom Method	Containment																																																						
Data Set	WORK.D1																																																																								
Response Variable	x																																																																								
Response Distribution	Binary																																																																								
Link Function	Logit																																																																								
Variance Function	Default																																																																								
Variance Matrix Blocked By	person																																																																								
Estimation Technique	Maximum Likelihood																																																																								
Likelihood Approximation	Gauss-Hermite Quadrature																																																																								
Degrees of Freedom Method	Containment																																																																								
Class Level Information																																																																									
<table border="1"> <tr><td>Class</td><td>Levels</td><td>Values</td></tr> <tr><td>item</td><td>5</td><td>1 2 3 4 5</td></tr> </table>		Class	Levels	Values	item	5	1 2 3 4 5																																																																		
Class	Levels	Values																																																																							
item	5	1 2 3 4 5																																																																							
Number of Observations Read 5000																																																																									
Number of Observations Used 5000																																																																									
:																																																																									
Dimensions																																																																									
<table border="1"> <tr><td>G-side Cov. Parameters</td><td>1</td></tr> <tr><td>Columns in X</td><td>6</td></tr> <tr><td>Columns in Z per Subject</td><td>1</td></tr> <tr><td>Subjects (Blocks in V)</td><td>1000</td></tr> <tr><td>Max Obs per Subject</td><td>5</td></tr> </table>		G-side Cov. Parameters	1	Columns in X	6	Columns in Z per Subject	1	Subjects (Blocks in V)	1000	Max Obs per Subject	5																																																														
G-side Cov. Parameters	1																																																																								
Columns in X	6																																																																								
Columns in Z per Subject	1																																																																								
Subjects (Blocks in V)	1000																																																																								
Max Obs per Subject	5																																																																								
Optimization Information																																																																									
<table border="1"> <tr><td>Optimization Technique</td><td>Dual Quasi-Newton</td></tr> <tr><td>Parameters in Optimization</td><td>6</td></tr> <tr><td>Lower Boundaries</td><td>1</td></tr> <tr><td>Upper Boundaries</td><td>0</td></tr> <tr><td>Fixed Effects</td><td>Not Profiled</td></tr> <tr><td>Starting From</td><td>GLM estimates</td></tr> <tr><td>Quadrature Points</td><td>5</td></tr> </table>		Optimization Technique	Dual Quasi-Newton	Parameters in Optimization	6	Lower Boundaries	1	Upper Boundaries	0	Fixed Effects	Not Profiled	Starting From	GLM estimates	Quadrature Points	5																																																										
Optimization Technique	Dual Quasi-Newton																																																																								
Parameters in Optimization	6																																																																								
Lower Boundaries	1																																																																								
Upper Boundaries	0																																																																								
Fixed Effects	Not Profiled																																																																								
Starting From	GLM estimates																																																																								
Quadrature Points	5																																																																								
Iteration History																																																																									
<table border="1"> <thead> <tr><th>Iteration</th><th>Restarts</th><th>Evaluations</th><th>Objective Function</th><th>Change</th><th>Max Gradient</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>4</td><td>5977.8672477</td><td>.</td><td>161.6788</td></tr> <tr><td>1</td><td>0</td><td>4</td><td>5908.882699</td><td>68.98454866</td><td>155.4896</td></tr> <tr><td>2</td><td>0</td><td>3</td><td>5903.1279635</td><td>5.75473554</td><td>125.9053</td></tr> <tr><td>3</td><td>0</td><td>2</td><td>5887.0364918</td><td>16.09147170</td><td>128.4802</td></tr> <tr><td>4</td><td>0</td><td>2</td><td>5868.1215222</td><td>18.91496956</td><td>22.19566</td></tr> <tr><td>5</td><td>0</td><td>3</td><td>5866.564838</td><td>1.55668425</td><td>16.59129</td></tr> <tr><td>6</td><td>0</td><td>3</td><td>5865.7471981</td><td>0.81763986</td><td>21.89317</td></tr> <tr><td>7</td><td>0</td><td>2</td><td>5865.1399141</td><td>0.60728395</td><td>6.7459</td></tr> <tr><td>8</td><td>0</td><td>3</td><td>5864.9314341</td><td>0.20648008</td><td>1.066567</td></tr> <tr><td>9</td><td>0</td><td>3</td><td>5864.9292926</td><td>0.00214149</td><td>0.112469</td></tr> <tr><td>10</td><td>0</td><td>3</td><td>5864.9292634</td><td>0.00002919</td><td>0.005109</td></tr> </tbody> </table>		Iteration	Restarts	Evaluations	Objective Function	Change	Max Gradient	0	0	4	5977.8672477	.	161.6788	1	0	4	5908.882699	68.98454866	155.4896	2	0	3	5903.1279635	5.75473554	125.9053	3	0	2	5887.0364918	16.09147170	128.4802	4	0	2	5868.1215222	18.91496956	22.19566	5	0	3	5866.564838	1.55668425	16.59129	6	0	3	5865.7471981	0.81763986	21.89317	7	0	2	5865.1399141	0.60728395	6.7459	8	0	3	5864.9314341	0.20648008	1.066567	9	0	3	5864.9292926	0.00214149	0.112469	10	0	3	5864.9292634	0.00002919	0.005109
Iteration	Restarts	Evaluations	Objective Function	Change	Max Gradient																																																																				
0	0	4	5977.8672477	.	161.6788																																																																				
1	0	4	5908.882699	68.98454866	155.4896																																																																				
2	0	3	5903.1279635	5.75473554	125.9053																																																																				
3	0	2	5887.0364918	16.09147170	128.4802																																																																				
4	0	2	5868.1215222	18.91496956	22.19566																																																																				
5	0	3	5866.564838	1.55668425	16.59129																																																																				
6	0	3	5865.7471981	0.81763986	21.89317																																																																				
7	0	2	5865.1399141	0.60728395	6.7459																																																																				
8	0	3	5864.9314341	0.20648008	1.066567																																																																				
9	0	3	5864.9292926	0.00214149	0.112469																																																																				
10	0	3	5864.9292634	0.00002919	0.005109																																																																				
Convergence criterion (GCONV=1E-8) satisfied.																																																																									
Fit Statistics																																																																									
<table border="1"> <tr><td>-2 Log Likelihood</td><td>5864.93</td></tr> <tr><td>AIC (smaller is better)</td><td>5876.93</td></tr> <tr><td>AICC (smaller is better)</td><td>5876.95</td></tr> <tr><td>BIC (smaller is better)</td><td>5906.38</td></tr> <tr><td>CAIC (smaller is better)</td><td>5912.38</td></tr> <tr><td>HQIC (smaller is better)</td><td>5888.12</td></tr> </table>		-2 Log Likelihood	5864.93	AIC (smaller is better)	5876.93	AICC (smaller is better)	5876.95	BIC (smaller is better)	5906.38	CAIC (smaller is better)	5912.38	HQIC (smaller is better)	5888.12																																																												
-2 Log Likelihood	5864.93																																																																								
AIC (smaller is better)	5876.93																																																																								
AICC (smaller is better)	5876.95																																																																								
BIC (smaller is better)	5906.38																																																																								
CAIC (smaller is better)	5912.38																																																																								
HQIC (smaller is better)	5888.12																																																																								
Fit Statistics for Conditional Distribution																																																																									
<table border="1"> <tr><td>-2 log L(x r. effects)</td><td>5509.09</td></tr> <tr><td>Pearson Chi-Square</td><td>4597.70</td></tr> <tr><td>Pearson Chi-Square / DF</td><td>0.92</td></tr> </table>		-2 log L(x r. effects)	5509.09	Pearson Chi-Square	4597.70	Pearson Chi-Square / DF	0.92																																																																		
-2 log L(x r. effects)	5509.09																																																																								
Pearson Chi-Square	4597.70																																																																								
Pearson Chi-Square / DF	0.92																																																																								
Covariance Parameter Estimates																																																																									
<table border="1"> <tr><th>Cov Parm</th><th>Subject</th><th>Estimate</th><th>Standard Error</th></tr> <tr><td>UN(1,1)</td><td>person</td><td>0.2110</td><td>0.06511</td></tr> </table>		Cov Parm	Subject	Estimate	Standard Error	UN(1,1)	person	0.2110	0.06511																																																																
Cov Parm	Subject	Estimate	Standard Error																																																																						
UN(1,1)	person	0.2110	0.06511																																																																						
:																																																																									
(continued)																																																																									

TABLE 13.2. (continued)

Solution for Random Effects						
Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	1	-0.09079	0.4185	3996	-0.22	0.8283
Intercept	2	0.08358	0.4180	3996	0.20	0.8415
Intercept	3	-0.2657	0.4241	3996	-0.63	0.5311
Intercept	4	0.08358	0.4180	3996	0.20	0.8415
Intercept	5	0.08358	0.4180	3996	0.20	0.8415
Intercept	6	0.2577	0.4225	3996	0.61	0.5419
•						
Intercept	1000	-0.2657	0.4241	3996	-0.63	0.5311
Item Least Squares Means						
item	Estimate	Standard Error	DF	t Value	Pr > t	Standard Error Mean
1	2.1303	0.1044	3996	20.40	< .0001	0.8938 0.00909
2	0.2745	0.06705	3996	4.09	< .0001	0.5682 0.01645
3	-0.04206	0.06644	3996	-0.63	0.5268 0.4895	0.01660
4	-0.8095	0.07185	3996	-11.27	< .0001	0.3080 0.01531
5	-1.0414	0.07537	3996	-13.82	< .0001	0.2609 0.01453
•						
The design matrix approach yields the following tables:						
•						
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error			
Intercept	person	0.2110	0.06511			
Solutions for Fixed Effects						
Effect	Estimate	Standard Error	DF	t Value	Pr > t	
Intercept	-1.0414	0.07537	999	-13.82	< .0001	
x1	3.1716	0.1297	3996	24.46	< .0001	
x2	1.3159	0.09930	3996	13.25	< .0001	
x3	0.9993	0.09827	3996	10.17	< .0001	
x4	0.2319	0.1006	3996	2.31	0.0212	
•						
Covariance Matrix for Fixed Effects						
Effect	Row	Col1	Col2	Col3	Col4	Col5
Intercept	1	0.005681	-0.00580	-0.00552	-0.00546	-0.00532
x1	2	-0.00580	0.01682	0.005942	0.005775	0.005386
x2	3	-0.00552	0.005942	0.009860	0.005511	0.005329
x3	4	-0.00546	0.005775	0.005511	0.009656	0.005317
x4	5	-0.00532	0.005386	0.005329	0.005317	0.01012

vergence. Our Fit Statistics table provides model-level fit information to be used for model comparisons.

Our person estimates are in the Solution for Random Effects table. The person estimates for the first couple of examinees show they have approximately average mathematics proficiency ($\hat{\theta}_1 = -0.09079$, $\hat{\theta}_2 = 0.08358$), the third examinee has below-average mathematics proficiency ($\hat{\theta}_1 = -0.2657$), and so on. As would be expected with the Rasch model, all individuals with the same observed score obtain the same $\hat{\theta}$. For example, examinees 2, 4, and 5 each have an estimated location of 0.08358 and an $X = 3$ ($\underline{x}'_2 = 11100$, $\underline{x}'_4 = 11100$, $\underline{x}'_5 = 10110$). The item location estimates are found in the Item Least Squares Means table. Our easiness values range from an easy item of $\hat{\delta}_1^E = 2.1303$ to a comparatively more difficult item of $\hat{\delta}_5^E = -1.0414$.^{12,13} We can transform these estimates to a traditional difficulty scale by multiplying by -1 either in a subsequent SAS data step or outside of SAS.

For pedagogical purposes, we present the use of the design matrix approach. Our item parameter estimates are found in the Solution for Fixed Effects table. Our reference item's estimated effect is given by the intercept, $\hat{\zeta}_{00}$, value of -1.0414 with item specific (fixed) estimated effects of 3.1716, 1.3159, 0.9993, and 0.2319 for items 1–4, respectively. We obtain our item location estimates (with respect to easiness) by Equation 13.13. As an example, for the first item $\hat{\delta}_1^E = \hat{\zeta}_{00} + \hat{\zeta}_{10} = -1.4011 + 3.1716 = 2.1302$. This value matches item 1's location from the Item Least Squares Means table. As can be seen, the standard errors for our fixed effect estimates do not match those seen in the Item Least Squares Means table. To correct our fixed effects standard errors, we need the covariances of each item with the intercept. We can obtain these covariances by using the covb option on the model line to produce the covariance matrix (see Table 13.1). To obtain the correct standard error for item j , we apply

$$s_e(\delta_j^E) = s_e(\delta_j) = \sqrt{s_e^2(\zeta_{00}) + s_e^2(\zeta_{j0}) + 2 \text{cov}(\zeta_{00}, \zeta_{j0})}, \quad (13.17)$$

where $s_e^2(\zeta_{00})$ and $s_e^2(\zeta_{j0})$ are the intercept's and the j th fixed effect's variance errors, respectively, and $\text{cov}(\zeta_{00}, \zeta_{j0})$ is their covariance (Roberts & Herrington, 2005). For example, for the first item we have $s_e(\hat{\delta}_1) = \sqrt{0.07537^2 + 0.1297^2 + 2(-0.00580)} = 0.104416$. This value matches the first item's standard error found in the Item Least Squares Means table.¹⁴

For comparison, we calibrate our data using proc irt (Table 13.3). The first data step in Table 13.1 shows the data step to read our data for this analysis (Figure 13.2 “Original (unstacked) Format” panel). The resulting data set, unstacked _ data, is used by proc irt (data = unstacked _ data). On our proc irt statement we request our traditional P-values, P_j , and item discriminations (ITEMSTAT), item fit information (i.e., Pearson's chi-square, G^2 ; itemfit), several plots (scree: eigenvalue plot, icc: item response function plot, iic: item information function, tic: test information function), to use 15 quadrature points for estimation (qpoints), and to use EAP for person estimation; ODS graphics must be turned on to obtain the plots. We output our person estimates to a data set (out = thetahat) that is displayed by using proc print (command not shown); item parameter estimates may also be outputted for manipulation/processing in subsequent data and/or proc steps. By default, the

TABLE 13.3. proc irt Program for Rasch Calibration^a

```
ods graphics on;

proc irt data=unstacked_data itemfit itemstat out=thetahat
plots=(scree icc iic tic)
scoremethod=eap qpoints=15 nfact=1 link=logit;
var i1-i5;
model i1-i5/resfunc=rasch;
run;
```

^aSee Table 13.1 for the corresponding data step.

`nfact` (number of factors) and `link` (link function) options are 1 and `logit`, respectively, but are included here to illustrate that `proc irt` can fit multidimensional as well as probit models (e.g., `nfact = 2, link = probit`). On the variable statement (`var`), we specify the items for calibration. Because `proc irt` allows different items to be calibrated using different models, the `model` statement identifies which items are to be calibrated using which model. In our example, all five items are calibrated using the Rasch model (`resfunc = rasch`). (Additional models include, for example, 1PL, 2PL, 3PL, 4PL, GR, GPC as well as models that contain parameter constraints.) Table 13.4 contains the corresponding output.

Inspection of the `Modeling Information` table shows the correct (1) response model (`Response Model Rasch Model`), (2) link function (`Link Function Logit`), and (3) number of cases. The `Item Information` table lists the correct number of binary items and response codes. Our `Item Statistics` table presents the traditional P-values (`Mean`), the point-biserial correlation (`Unadjusted Item-Total Correlation`), and the corrected point-biserial correlation (`Adjusted Item-Total Correlation`). (The adjustment of an item's point-biserial mitigates the spuriously high item-total (score) correlation due to the effect of the item being included in the total (i.e., observed) score.) When there are 40 or more items, the difference between the adjusted and unadjusted item-total correlations may not be of practical significance. Not surprisingly, with five items the adjustment is substantial. As can be seen, there is some variability in these discrimination indices corresponding to the variability in our P-values (i.e., `Mean`) with moderate item difficulty indices exhibiting the best discrimination. Moreover, our data are divided into four fractiles (`G1, G2, G3, G4`) of “approximately equal” size, with the fractile mean for each item displayed. For a given item, we expect to see the proportion correct (i.e., `mean`) to increase as we move from a “lower ability” group (e.g., `G1`) to a “higher ability” group (e.g., `G2`).

Although not shown, SAS will present a table of eigenvalues along with the corresponding scree plot to aid in assessing dimensionality. The correlation matrix analyzed is a polychoric correlation matrix. As such, with binary data the polychoric correlation matrix simplifies to a tetrachoric matrix. Our iteration history is presented next. Because we obtained convergence (in ten iterations), we examine the remaining output.

As discussed above, our model-level fit statistics (`Model Fit Statistics`) can be used to compare our results with those of other models. At the item level, the significant Pearson chi-square and Likelihood Ratio (`LR chi-square`) statistic indicate that none of the items are consistent with the Rasch model.¹⁵ However, the small number of items makes the calculation of these significance tests problematic.¹⁶ As a result, we ignore these item-level statistical tests.

Our item calibration results follow in the `Item Parameter Estimates` table. As can be seen, our location estimates (with respect to difficulty) are $\hat{\delta}_1 = -2.13029$, $\hat{\delta}_2 = -0.27454$, $\hat{\delta}_3 = 0.04206$, $\hat{\delta}_4 = 0.80949$, and $\hat{\delta}_5 = 1.04135$. (The probability displayed $\Pr > |t|$) is associated with a test of the null hypothesis that the parameter is equal to 0.) The corresponding IRFs are presented next with “`x =`” indicating the location estimates value and the vertical line indicating the item's location on the scale. Because this model's lower asymptote is zero, the corresponding `p` at the item's location is 0.5

TABLE 13.4. proc irt Output^a

(continued)

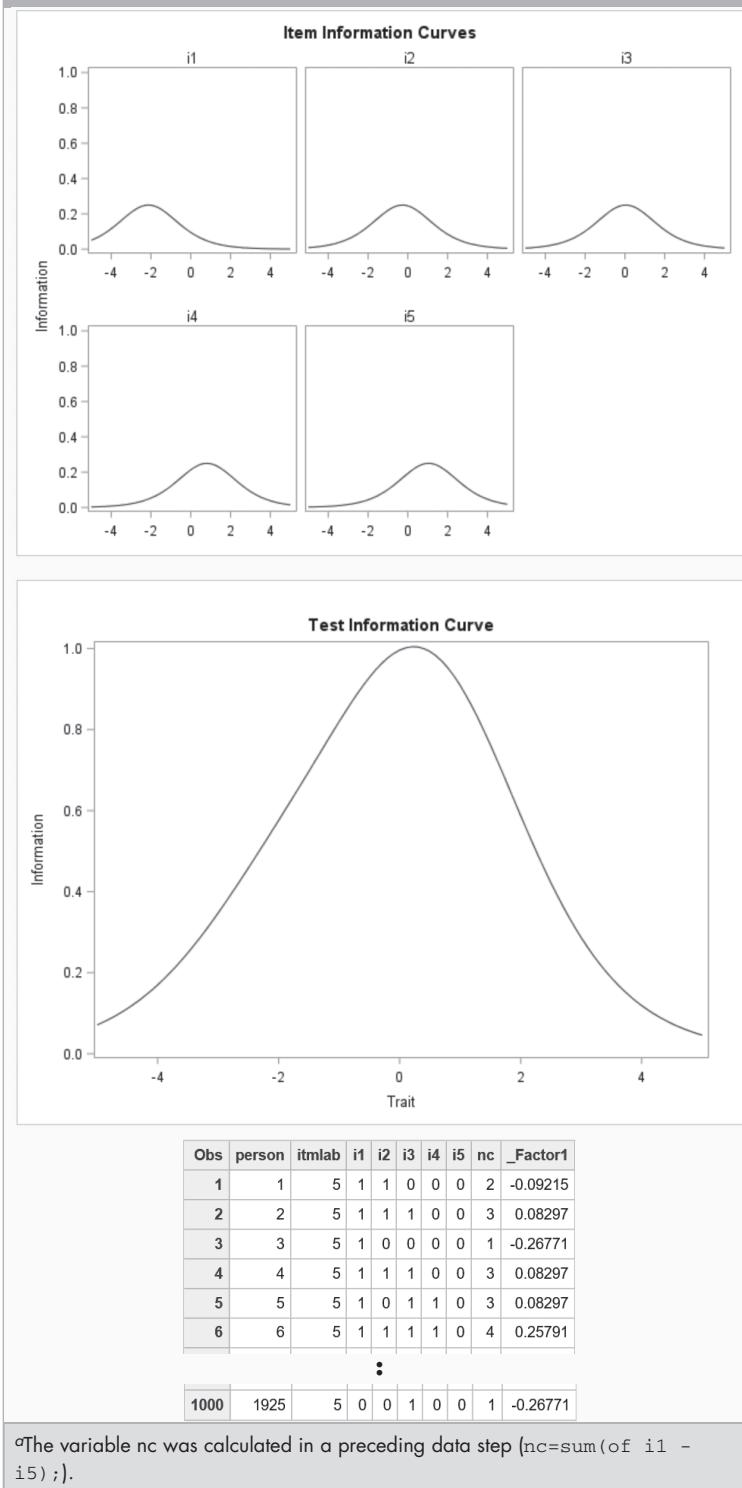
TABLE 13.4. (continued)

The IRT Procedure						
Model Fit Statistics						
						Log Likelihood
						-2932.464664
						AIC (Smaller is Better)
						5876.9293279
						BIC (Smaller is Better)
						5906.3758596
						LR Chi-Square
						255.31665491
						LR Chi-Square DF
						25
Item Fit Statistics						
Item	DF	Pearson Chi-Square	Pr > P ChiSq	LR Chi-Square	Pr > LR ChiSq	
i1	9	82.29790	<.0001	93.76968	<.0001	
i2	9	622.49168	<.0001	826.41844	<.0001	
i3	9	368.99898	<.0001	474.31405	<.0001	
i4	9	164.91621	<.0001	177.69978	<.0001	
i5	9	307.15087	<.0001	381.99761	<.0001	

The IRT Procedure						
Item Parameter Estimates						
Item	Parameter	Estimate	Standard Error	Pr > t		
i1	Difficulty	-2.13029	0.10440	<.0001		
	Slope	1.00000				
i2	Difficulty	-0.27454	0.06705	<.0001		
	Slope	1.00000				
i3	Difficulty	0.04206	0.06644	0.2634		
	Slope	1.00000				
i4	Difficulty	0.80949	0.07185	<.0001		
	Slope	1.00000				
i5	Difficulty	1.04135	0.07537	<.0001		
	Slope	1.00000				

Item Characteristic Curves						
i1	x = -2.1	0.00	0.25	0.50	0.75	1.00
i2	x = -.27	0.00	0.25	0.50	0.75	1.00
i3	x = 0.04	0.00	0.25	0.50	0.75	1.00
i4	x = 0.81	0.00	0.25	0.50	0.75	1.00
i5	x = 1.04	0.00	0.25	0.50	0.75	1.00
Item Characteristic Curves						
:						

(continued)

TABLE 13.4. (continued)

(i.e., the horizontal line). As would be expected, each item provides the same maximum information albeit at different locations throughout the continuum (see Item Information Curves plots). The Test Information Curve presents the instrument's total information and shows the instrument provides the most information in a neighborhood around 0.

We display our person location estimates by using `proc print`. Thus, person 1 ($\bar{x}_1 = 11000$) obtained an $X_1 = 2$ and is estimated to be located at $(\hat{\theta} =) -0.09215$; persons 2, 4, and 5 with $X = 3$ are estimated to be located at $\hat{\theta} = 0.0829$; and so on.¹⁷

Comparing the `proc glimmix` and `proc irt` item location estimates shows perfect agreement with one another in terms of magnitude. As mentioned above, `glimmix` estimates item locations in terms of their easiness, whereas `proc irt` estimates item locations with respect to their difficulty. Easiness and difficulty are, in effect, a mirror image of one another. Similarly, we have perfect agreement between the two sets of person location estimates with a correlation of 1.00000. On the latent continuum, the `glimmix` person estimates are almost perfectly aligned with the conventional IRT approach. Descriptive statistics for `glimmix` are an average $\hat{\delta}$ of 0.10237 ($SD = 1.12274$) with a mean $\hat{\theta}$ of 0.00093 ($SD = 0.18841$), whereas for `proc irt` the mean $\hat{\delta} = -0.10239$ ($SD = 1.12272$) and the average $\hat{\theta}$ is 0.00000 ($SD = 0.18921$). This example and the process of obtaining Equation 13.16 demonstrate the equivalency of the Rasch model and its multilevel formulation.¹⁸

Example: Rasch Model Estimation, lme4

For our multilevel estimation, we use the `lme4` (Bates, Mächler, Bolker, & Walker, 2015) package that estimates linear, generalized linear, and nonlinear mixed models. Our R session (Table 13.5) begins with loading `lme4` and two utility packages. The first utility package, `tidyverse` (Wickham, 2017), includes `tidyr` (Wickham, 2019) whose `gather` function we use to convert our unstacked data to a stacked format. Our second utility package (`optimx`) is an optimization package.

We read our data file (`mathunstacked.dat`) using the `file.choose()` function to interactively select the file from a dialog window and assign it to a data frame called `unstacked`. To transform our unstacked data into the stacked format shown in Figure 13.2 requires two steps. In our first step, we use the `gather` function to convert our unstacked data to stacked data. In the call to the `gather` function, we specify `item` as the name for the variable to contain the item labels, `x` to be the variable name for the corresponding responses, followed by the variables we wish to stack `i1:i5` where `i1:i5` is shorthand for `i1, i2, i3, i4`, and `i5`. The result of the `gather` function is the creation of the `tmpstacked` data frame whose format has all of the responses to item 1 on lines 1–1000, followed by all of the responses to item 2 on lines 1001–2000, followed by all of the responses to item 3 on lines 2001–3000, and so on. Thus, the first examinee's item responses appear on lines 1, 1001, 2001, 3001, and 4001, and so on for the other examinees. Our second step rearranges (`arrange`) the stacked data so that all the item responses for an examinee are contiguous.¹⁹

TABLE 13.5. lme4 Session for Rasch Calibration

```

> library(lme4)
> packageVersion("lme4")
[1] '1.1.21'
> library(tidyverse)
> packageVersion("tidyverse")
[1] '1.2.1'
> library(optimx)
> packageVersion("optimx")
[1] '2018.7.10'

> # data file: mathunstacked.dat
> unstacked = read.table(file.choose(), col.names=c("person",paste0("i",1:5)))

> head(unstacked,5)
  person i1 i2 i3 i4 i5
1      1  1  1  0  0  0
2      2  1  1  1  0  0
3      3  1  0  0  0  0
4      4  1  1  1  0  0
5      5  1  0  1  1  0

> tail(unstacked,5)
  person i1 i2 i3 i4 i5
996    1920  1  1  0  1  0
997    1921  1  0  0  1  0
998    1922  0  1  0  0  1
999    1924  1  1  1  0  0
1000   1925  0  0  1  0  0

> # reformat unstacked data into stacked format (tidyr: gather)
> tmpstacked = gather(unstacked, key=item,value=x,i1:i5)

> # and reorder data to be items w/i person (tidyverse: arrange)
> stacked= arrange(tmpstacked, person,item)

> head(stacked,20)
  person item x
1      1   i1  1
2      1   i2  1
3      1   i3  0
4      1   i4  0
5      1   i5  0
6      2   i1  1
7      2   i2  1
8      2   i3  1
9      2   i4  0
10     2   i5  0
11     3   i1  1
12     3   i2  0
13     3   i3  0
14     3   i4  0
15     3   i5  0
16     4   i1  1
17     4   i2  1
18     4   i3  1
19     4   i4  0
20     4   i5  0

```

(continued)

TABLE 13.5. (*continued*)

```

># to speed up execution
> optmzrinfo=glmerControl(optimizer="optimx",calc.derivs=F,optCtrl=list
+                               (method = "nlsinb",starttests=F,kkt=F))

> # doing the rasch model w/o intercept
> rasch=glmer(x~0+item+(1|person),family=binomial("logit"),data=stacked,
+               control=optmzrinfo)

> summary(rasch)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + item + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC      logLik deviance df.resid
5878.8   5917.9   -2933.4    5866.8     4994

Scaled residuals:
    Min      1Q  Median      3Q      Max
-3.2323 -0.6933  0.3094  0.8407  1.8874

Random effects:
Groups Name        Variance Std.Dev.
person (Intercept) 0.1811   0.4256
Number of obs: 5000, groups: person, 1000

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
itemi1    2.11963   0.10285 20.609 < 2e-16 ***
itemi2    0.27351   0.06545  4.179  2.93e-05 ***
itemi3   -0.04184   0.06489 -0.645   0.519
itemi4   -0.80598   0.06989 -11.533 < 2e-16 ***
itemi5   -1.03666   0.07325 -14.152 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
       itemi1 itemi2 itemi3 itemi4
itemi2  0.027
itemi3  0.027  0.043
itemi4  0.025  0.040  0.040
itemi5  0.024  0.038  0.038  0.036

> thetahat = coef(rasch)$person[,1]    # extracting respondent est
> head(thetahat,21)
[1] -0.0799455  0.0734883 -0.2337150  0.0734883  0.0734883  0.2267480  0.2267480
[8] -0.0799455 -0.2337150 -0.2337150 -0.0799455 -0.2337150  0.0734883  0.0734883
[15] -0.0799455  0.2267480 -0.0799455 -0.0799455 -0.2337150  0.0734883 -0.0799455
> peopleRasch=matrix(thetahat)
> peopleRasch
      [,1]
[1,] -0.0799455
[2,]  0.0734883
[3,] -0.2337150
[4,]  0.0734883
[5,]  0.0734883
[6,]  0.2267480
:
[1000,] -0.2337150

```

Multilevel modeling can be computationally intensive and time consuming. Thus, to minimize execution time, we use the wrapper function `optimx` (Nash & Varadhan, 2011) with the `nlminb` optimizing method. For pedagogical reasons we separate the command for optimization (`optmzrinfo = glmerControl(optimizer ...)`) from the command for performing the parameter estimation (`rasch = glmer(x..., control = optmzrinfo)`).

We invoke the optimizer through `glmerControl` specifying the `nlminb` method and not to perform specific initial processing (e.g., not running scaling tests before optimizing nor the Kuhn, Karush, Tucker optimality conditions); for greater details see <https://www.rdocumentation.org/packages/optimx/versions/2018-7.10/topics/optimx>.²⁰ The `lme4` syntax uses parentheses to indicate random effects and nonparenthesized terms to reflect fixed effects as well as the numbers 0 (or -1) and 1 to indicate the absence or presence of an intercept, respectively; if the intercept is to be treated as random, the "1" would be placed within the parenthesized term. In general, our model has structure:

<outcome variable> ~ (#) + fixed effect(s) + (random effect | nesting variable) . . .

Additionally and as we do in `glimmix`, we need to specify the response's distribution (`binomial`) and the link function.

We use `lme4`'s `glmer` ("generalized linear mixed-effects in R") function and assign the results to the `rasch` object; `lme4` can also fit nonlinear mixed-effects models (`nlmer`) and negative binomial GLMMs (`glmer.nb`). Our random outcome variable, `x`, is separated from the linear component by a tilde; see De Boeck et al. (2011). Therefore, `x ~ 0 + item + (1|person)` indicates that our response variable `x` is to be modeled by a no-intercept (0) model treating `item` as fixed (constant) across our examinees with random intercepts across our examinees ((1|person)); see Endnote 14. Because we are modeling a dichotomous response variable, we specify our response distribution to be `binomial`. For pedagogical reasons, we specify the default logistic link, `logit`. We also indicate the use of the stacked data (frame) and our optimization instructions (`control = optmzrinfo`).

Our item easiness estimates are found in the Estimate column in the Fixed effects: section of our output; $\hat{\delta}_1^E = 2.11963$, $\hat{\delta}_2^E = 0.27351$, . . . , $\hat{\delta}_5^E = -1.03666$. These values agree with those of `glimmix` and `proc irt` and show a correlation with the `glimmix` and `proc irt` estimates of 1.0 and -1.0, respectively. When we align the `glmer` estimates' metric ($M = 0.101732$, $SD = 1.11731$) with those of `glimmix` and `proc irt`, our estimates show an average discrepancy from the latter two of 0.00000.

We abstract our person estimates with `thetahat = coef(rasch)$person[,1]` and then reformat them into a single column using `matrix`. As can be seen, person 1 ($X_1 = 2$) is estimated to be located at $(\hat{\theta} =) -0.0799455$, persons 2, 4, and 5 ($X = 3$) are estimated to be located at $\hat{\theta} = 0.0734883$, and so on. Comparing these estimates to those of `glimmix` shows that they are only slightly different due to `glmer`'s and `glimmix`'s metrics being slightly unaligned. The mean `glmer` person estimate ($M = 0.000799$, $SD = 0.16576$) is slightly different from `glimmix`'s average person esti-

mate ($M = 0.00093$, $SD = 0.18841$). When we take into account the differences in means and standard deviations, the two sets of person location estimates are identical; the same is true with proc irt.

Person-Level Predictors for Items

We now discuss the use of person-oriented predictors with our items. As mentioned above, we can include additional predictors in our level-2 model to account for the observed variability in the β s depicted in Figure 13.1. For example, we can add person-level predictors to assess the impact of person characteristics on our item parameter estimates. These person-level characteristics could reflect different categorical groupings such as gender, race, and ethnicity. As such, we can use multilevel IRT models to perform a DIF analysis by including a grouping predictor in our person-level model; our item-level model (i.e., Equation 13.7) remains the same.

As an example, assume we wish to examine whether items are more difficult for one group versus another (e.g., females vs. males). To model this possibility, our (level-2) person-level models would include a dichotomous group predictor (e.g., *Group*) of our level-1 item effects (i.e., regression coefficients). Following Kamata (1998, 2001; also see Cheong & Kamata, 2013), we have at level 2

$$\begin{aligned}\beta_{1i} &= \varsigma_{10} + \varsigma_{11}Group_i \\ \beta_{2i} &= \varsigma_{21} + \varsigma_{20}Group_i \\ &\vdots \\ \beta_{(L-1)i} &= \varsigma_{(L-1)0} + \varsigma_{(L-1)1}Group_i.\end{aligned}\tag{13.18}$$

In Equation 13.18 the β_{qi} regression coefficients are allowed to vary across our groups with the ς_{qj} s indicating the magnitude of any group difference in the item effect j ($q = j$) and the indicator predictor $Group_i$ having a code of 0 if person i is a Reference group member or a code of 1 otherwise. The constants ς_{q0} is the item effect for the Reference group members.

We divide our items into reference and nonreference items. For our reference item, person i 's log odds for a response of 1 on the item is

$$\beta_{0i} = \varsigma_{00} + \varsigma_{01}Group_i + \theta_{0i}^r,\tag{13.19}$$

where θ_{0i}^r is person i 's location estimate adjusted for group membership (see Kamata, 2001; Cheong & Kamata, 2013). Our reference item constant ς_{00} is the average performance for the Reference group members (i.e., the group coded 0), with the coefficient ς_{01} reflecting the average difference in log odds of a response of 1 between Reference

and Focal group members. The “reference item is assumed to display no DIF” (Cheong & Kamata, 2013, p. 239).

The log odds of a response of 1 to item j ($q = j$) is given by the combined model

$$\begin{aligned}
 \vartheta_{ji} &= \beta_{0i} + \beta_{qi} \\
 &= [\varsigma_{00} + \varsigma_{01}Group_i + \theta_{0i}^r] - [\varsigma_{q0} + \varsigma_{q1}Group_i] \\
 &= \varsigma_{00} + \varsigma_{01}Group_i + \theta_{0i}^r - \varsigma_{q0} - \varsigma_{q1}Group_i \\
 &= \theta_{0i}^r + \varsigma_{00} - \varsigma_{q0} - (\varsigma_{q1} - \varsigma_{01})Group_i \\
 &= \theta_{0i}^r - [\varsigma_{q0} - \varsigma_{00} + (\varsigma_{q1} - \varsigma_{01})Group_i]. \tag{13.20}
 \end{aligned}$$

Distinguishing between our $(L - 1)$ nonreference items and our reference item (i.e., the L th item), we have as the location of a nonreference item

$$\delta_j = [\varsigma_{q0} - \varsigma_{00} + (\varsigma_{q1} - \varsigma_{01})Group_i], \tag{13.21}$$

with the effect of *Group* given by $(\varsigma_{q0} - \varsigma_{01})$.

For the reference item (i.e., “ $q = L$ ”) we have, by definition, that $\varsigma_{q0} = \varsigma_{q1} = 0$, and thus its location is

$$\delta_L = -\varsigma_{00} - \varsigma_{01}Group_i, \tag{13.22}$$

with *Group*’s effect given by ς_{01} . We can use Equations 13.21 and 13.22 to identify the absence (or presence) of DIF.

For our $(L - 1)$ nonreference items we substitute the appropriate Focal and Reference codes into the $Group_i$ predictor. For instance, for a nonreference item location and given Equation 13.21 we have for a Focal group member ($Group_i = 1$)

$$\delta_{jF} = \varsigma_{q0} - \varsigma_{00} + (\varsigma_{q1} - \varsigma_{01})(1) = (\varsigma_{q0} - \varsigma_{00}) + (\varsigma_{q1} - \varsigma_{01})$$

and for a Reference group member ($Group_i = 0$)

$$\delta_{jR} = \varsigma_{q0} - \varsigma_{00} + (\varsigma_{q1} - \varsigma_{01})(0) = (\varsigma_{q0} - \varsigma_{00}).$$

Therefore, there is no DIF when the last term for the Focal group, $(\varsigma_{q1} - \varsigma_{01})$, equals 0 so that we have $\delta_{jF} = \delta_{jR}$.

With respect to the reference item and given Equation 13.22, we have for a member of the Focal group ($Group_i = 1$)

$$\delta_{LF} = -\varsigma_{00} - \varsigma_{01}(1) = -\varsigma_{00} - \varsigma_{01}$$

and for a Reference group member ($Group_i = 0$)

$$\delta_{LR} = -\zeta_{00} - \zeta_{01}(0) = -\zeta_{00}.$$

When the last term for the Focal group, ζ_{01} , equals 0 we have $\delta_{LF} = \delta_{LR}$ (i.e., no DIF). However, whenever $(\zeta_{q1} - \zeta_{01})$ is significantly different from 0, then $\delta_{jF} \neq \delta_{jR}$ and we have uniform DIF for item j . Similarly, if ζ_{01} is significantly different from 0, then $\delta_{LF} \neq \delta_{LR}$. The null hypotheses evaluated by these statistical tests are $H_0: (\zeta_{q1} - \zeta_{01}) = 0$ for the nonreference items and $H_0: \zeta_{01} = 0$ for the reference item. For additional information on DIF analyses from a multilevel see Cho and Cohen (2010), French and Finch (2010), Swanson, Clauser, Case, Nungester, and Featherman (2002), as well as Van den Noortgate and De Boeck (2005).

Example: Person-Level Predictors for Items— DIF Analysis, proc glimmix

We revisit our data from Chapter 12. Recall these data arose from the administration of a 50-item four-option multiple-choice vocabulary test. The test consisted of 40 items drawn from the verbal section of the College Qualification Test and involved standard English vocabulary and 10 “Black slang” (B) items. Each item presented a word whose definition the examinee chose from one of the four options. As in Chapter 12, we use the first 10 items with items 3 and 7 as B items. Our race indicator predictor is coded with the Reference group (Caucasian students) coded 0 and the Focal group (African Americans students) coded 1.

With the multilevel approach, all items are examined for DIF simultaneously. Table 13.6 contains our `glimmix` syntax for the analyses with and without our examinees’ race information. (The data steps from Table 13.1 (not shown) are used to create the

TABLE 13.6. proc glimmix Programs for DIF Analysis

```
proc glimmix data= stacked_data noclprint method=quadrature;
  title "no DIF model: no RACE predictor model";
  class item race;
  model x(desc)= item / solution dist=binary link=logit;
  random intercept / subject=person;
  lsmeans item / ilink;
run;

proc glimmix data= stacked_data noclprint method=quadrature;
  title "DIF model: RACE predictor model";
  class item race;
  model x(desc)= item race item*race / solution dist=binary link=logit;
  random intercept / subject=person;
  lsmeans item*race / ilink diff=all;
run;
```

stacked data set from the file vocabhdr.dat.) On each of the proc statements, we specify noclprint to suppress the printing of class-level identification. For comparison with our DIF results, our first proc glimmix simply calibrates the data according to the Rasch model. The second proc glimmix is for our DIF analysis. For the DIF analysis, our model line contains the main effects of item and race and their interaction (item*race). Additionally, we request that all pairwise differences between the least squares means be provided (diff = all).

Tables 13.7 and 13.8 contain abridged results for our no DIF and DIF analyses, respectively. Contrasting the no DIF analysis fit statistics with those from DIF analysis shows that, in terms of the deviance statistics, the model–data fit of the DIF model is significantly better than the no DIF model (i.e., $G_{\text{null}}^2 = 21,262.05$ and $G_{\text{DIF}}^2 = 20,000.83$ for a difference of 1261.22 on 10 df). Furthermore, all the information criteria favor the DIF model (e.g., $\text{BIC}_{\text{null}} = 21,345.82$ and $\text{BIC}_{\text{DIF}} = 20,160.76$).

TABLE 13.7. proc glimmix Output (No DIF Model)

Fit Statistics							
-2 Log Likelihood			21262.05				
AIC (smaller is better)			21284.05				
AICC (smaller is better)			21284.06				
BIC (smaller is better)			21345.82				
CAIC (smaller is better)			21356.82				
HQIC (smaller is better)			21306.71				
Covariance Parameter Estimates							
Cov Parm	Subject	Estimate	Standard Error				
Intercept	person	0.6093	0.04401				
item Least Squares Means							
item	Estimate	Standard Error	DF	t Value	Pr > t	Mean	Standard Error Mean
1	1.1112	0.05593	18260	19.87	<.0001	0.7523	0.01042
2	0.5918	0.05195	18260	11.39	<.0001	0.6438	0.01191
3	0.8165	0.05336	18260	15.30	<.0001	0.6935	0.01134
4	0.2067	0.05059	18260	4.09	<.0001	0.5515	0.01251
5	2.4223	0.07921	18260	30.58	<.0001	0.9185	0.005929
6	1.2535	0.05748	18260	21.81	<.0001	0.7779	0.009931
7	-2.3662	0.07780	18260	-30.41	<.0001	0.08579	0.006102
8	2.2132	0.07395	18260	29.93	<.0001	0.9014	0.006571
9	-1.3400	0.05853	18260	-22.89	<.0001	0.2075	0.009625
10	0.5633	0.05181	18260	10.87	<.0001	0.6372	0.01198

TABLE 13.8. proc glimmix Output (DIF Model)

Fit Statistics										
-2 Log Likelihood						20000.83				
AIC (smaller is better)						20042.83				
AICC (smaller is better)						20042.88				
BIC (smaller is better)						20160.76				
CAIC (smaller is better)						20181.76				
HQIC (smaller is better)						20086.10				
:										
Covariance Parameter Estimates										
Cov Parm	Subject	Estimate	Standard Error							
UN(1,1)	person	0.5627	0.04416							
:										
Solutions for Fixed Effects										
Effect	item	race	Estimate	Standard Error	DF	t Value	Pr > t			
Intercept			-0.1638	0.07093	2028	-2.31	0.0210			
item	1		0.6147	0.09539	18252	6.44	<.0001			
item	2		0.004629	0.09453	18252	0.05	0.9609			
item	3		1.9942	0.1121	18252	17.79	<.0001			
item	4		-0.2714	0.09527	18252	-2.85	0.0044			
item	5		1.9017	0.1101	18252	17.26	<.0001			
item	6		1.0085	0.09784	18252	10.31	<.0001			
item	7		-1.9527	0.1189	18252	-16.42	<.0001			
item	8		1.7166	0.1067	18252	16.09	<.0001			
item	9		-1.3353	0.1057	18252	-12.64	<.0001			
item	10		0			
race		0	1.5670	0.1093	18252	14.33	<.0001			
race		1	0			
item*race	1	0	-0.04515	0.1544	18252	-0.29	0.7700			
item*race	1	1	0			
item*race	2	0	0.06549	0.1470	18252	0.45	0.6560			
item*race	2	1	0			

(continued)

As can be seen from the Covariance Parameter Estimates tables, the inclusion of race information reduced our person variance from 0.6093 to 0.5629. The Solutions for Fixed Effects table (Table 13.8) shows, relative to that of being in the Focal group member, the effect of being a Reference group member is on average an increase of 1.5670.

TABLE 13.8. (*continued*)

item*race Least Squares Means									
item	race	Estimate	Standard Error	DF	t Value	Pr > t	Mean	Standard Error Mean	
1	0	1.9727	0.09635	18252	20.47	<.0001	0.8779	0.01033	
1	1	0.4509	0.07199	18252	6.26	<.0001	0.6108	0.01711	
2	0	1.4733	0.08444	18252	17.45	<.0001	0.8136	0.01281	
2	1	-0.1592	0.07092	18252	-2.24	0.0248	0.4603	0.01762	
3	0	0.03393	0.07060	18252	0.48	0.6309	0.5085	0.01764	
3	1	1.8304	0.09287	18252	19.71	<.0001	0.8618	0.01106	
4	0	0.8739	0.07535	18252	11.60	<.0001	0.7056	0.01565	
4	1	-0.4352	0.07193	18252	-6.05	<.0001	0.3929	0.01716	
5	0	3.7403	0.1873	18252	19.97	<.0001	0.9768	0.004243	
5	1	1.7379	0.09052	18252	19.20	<.0001	0.8504	0.01152	
6	0	1.7201	0.08978	18252	19.16	<.0001	0.8481	0.01156	
6	1	0.8447	0.07516	18252	11.24	<.0001	0.6994	0.01580	
7	0	-2.6203	0.1192	18252	-21.99	<.0001	0.06784	0.007537	
7	1	-2.1165	0.1013	18252	-20.90	<.0001	0.1075	0.009716	
8	0	3.3652	0.1595	18252	21.10	<.0001	0.9666	0.005149	
8	1	1.5528	0.08630	18252	17.99	<.0001	0.8253	0.01244	
9	0	-1.1736	0.07915	18252	-14.83	<.0001	0.2362	0.01428	
9	1	-1.4991	0.08527	18252	-17.58	<.0001	0.1826	0.01272	
10	0	1.4032	0.08311	18252	16.88	<.0001	0.8027	0.01316	
10	1	-0.1638	0.07093	18252	-2.31	0.0209	0.4591	0.01761	

(continued)

Our Item*race Least Squares Means table (Table 13.8) shows that, for example, item 1 is substantially easier for the Reference group ($\hat{\delta}_{1,R}^E = 1.9727$) than for the Focal group ($\hat{\delta}_{1,F}^E = 0.4509$). In contrast, for our first B item Focal group, members found the item to be easier ($\hat{\delta}_{3,F}^E = 1.8304$) than those from the Reference group ($\hat{\delta}_{3,R}^E = 0.03393$). In the Differences of Item*race Least Squares Means table, we find all pairwise differences between item estimates and their significance; not all pairwise differences are of interest (e.g., item 1 Race = 0 and item 2 Race = 0). For item 1, the difference between the Reference and Focal groups' easiness estimates ($\hat{\delta}_{1,Diff}^E = 1.9727 - 0.4509 = 1.5218$) is significant $\alpha = 0.05$ with a *t*-value of 12.68 at (line: item = 1, race = 0, _ item = 1, _ race = 1). Therefore, item 1 is performing differentially across racial groups and is favoring the Reference group. In contrast, item 3 also shows significant differential performance ($t = -15.40$) with the *t*'s sign indicating that the item favors the Focal group ($\hat{\delta}_{3,Diff}^E = 0.03393 - 1.8304 = -1.7965$); line: item = 3, race = 0, _ item = 3, _ race = 1).²¹

TABLE 13.8. (continued)

Differences of item*race Least Squares Means									
item	race	_item	_race	Estimate	Standard Error	DF	t Value	Pr > t	
1	0	1	1	1.5218	0.1200	18252	12.68	<.0001	
1	0	2	0	0.4994	0.1223	18252	4.08	<.0001	
1	0	2	1	2.1319	0.1197	18252	17.81	<.0001	
1	0	3	0	1.9388	0.1147	18252	16.90	<.0001	
1	0	3	1	0.1423	0.1329	18252	1.07	0.2843	
:									
2	0	2	1	1.6325	0.1103	18252	14.79	<.0001	
2	0	3	0	1.4394	0.1049	18252	13.72	<.0001	
2	0	3	1	-0.3571	0.1247	18252	-2.86	0.0042	
2	0	4	0	0.5994	0.1074	18252	5.58	<.0001	
:									
3	0	3	1	-1.7965	0.1166	18252	-15.40	<.0001	
3	0	4	0	-0.8400	0.09773	18252	-8.60	<.0001	
3	0	4	1	0.4691	0.1008	18252	4.65	<.0001	
3	0	5	0	-3.7064	0.1974	18252	-18.78	<.0001	
:									
7	0	7	1	-0.5039	0.1553	18252	-3.24	0.0012	
7	0	8	0	-5.9855	0.1981	18252	-30.21	<.0001	
7	0	8	1	-4.1731	0.1481	18252	-28.19	<.0001	
7	0	9	0	-1.4467	0.1379	18252	-10.49	<.0001	
:									
9	0	9	1	0.3255	0.1157	18252	2.81	0.0049	
9	0	10	0	-2.5768	0.1108	18252	-23.25	<.0001	
9	0	10	1	-1.0098	0.1062	18252	-9.51	<.0001	
9	1	10	0	-2.9023	0.1197	18252	-24.24	<.0001	
9	1	10	1	-1.3353	0.1057	18252	-12.64	<.0001	
10	0	10	1	1.5670	0.1093	18252	14.33	<.0001	

Example: Person-Level Predictors for Items—DIF Analysis, lme4

Our lme4 analysis (Table 13.9) parallels our Rasch calibration. After reading in our unstacked data (vocabhdr.dat) and reformatting it to be stacked, we first perform a Rasch calibration using the glmer function with the results sent to the rasch output object: `rasch = glmer(x~0 + item + (1|person), ...)`. Our information criterion BIC is 21,385.0 with a G^2_{null} of 21,275.9. Despite the use of two different estimation algorithms in glmer and glimmix (i.e., Laplace and adaptive quadrature), the glim-

TABLE 13.9. lme4 Session for DIF Analysis

```

> library(tidyverse); library(lme4); library(optimx)

> # data: vocabhdr.dat
> unstacked = read.table(file.choose(), header=T)

> # race = 0: Caucasian; race = 1: AfrAm

> head(unstacked, 5)
  person race i01 i02 i03 i04 i05 i06 i07 i08 i09 i10
  1      1   0   1   1   1   0   1   1   1   1   1   0
  2      2   0   0   1   1   1   1   0   0   1   0   1
  3      3   0   1   1   0   0   1   1   0   1   0   1
  4      4   0   1   0   0   0   1   1   0   1   0   1
  5      5   0   1   1   1   1   1   1   0   1   1   1

> # reformat unstacked data into stacked format & reorder data to be items
  w/i person
> tmpstacked=gather(unstacked, key=item, value=x, i01:i10)

> head(tmpstacked, 10)
  person race item x
  1      1   0   i01 1
  2      2   0   i01 0
  3      3   0   i01 1
  4      4   0   i01 1
  5      5   0   i01 1
  6      6   0   i01 1
  7      7   0   i01 0
  8      8   0   i01 1
  9      9   0   i01 1
  10     10  0   i01 1

> stacked=arrange(tmpstacked, person, race, item)
> head(stacked, 10)
  person race item x
  1      1   0   i01 1
  2      1   0   i02 1
  3      1   0   i03 1
  4      1   0   i04 0
  5      1   0   i05 1
  6      1   0   i06 1
  7      1   0   i07 1
  8      1   0   i08 1
  9      1   0   i09 1
  10     1   0   i10 0

># to speed up execution optmzrinfo - see Table 13.5
> stacked$race=as.factor(stacked$race)          # convert numeric values to be factor levels

> # rasch model w/o intercept
> rasch=glmer(x~0+item+(1|person), family=binomial("logit"), data=stacked,
  control=optmzrinfo)

> summary(rasch)
  Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]

```

(continued)

TABLE 13.9. (continued)

```

Family: binomial ( logit )
Formula: x ~ 0 + item + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC   logLik deviance df.resid
21297.9  21385.0 -10637.9  21275.9     20289

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.6999 -0.5790  0.3399  0.5868  5.8729

Random effects:
Groups Name        Variance Std.Dev.
person (Intercept) 0.5877   0.7666
Number of obs: 20300, groups: person, 2030

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
itemi01  1.11138   0.05461 20.352 < 2e-16 ***
itemi02  0.59220   0.05047 11.733 < 2e-16 ***
itemi03  0.81697   0.05193 15.733 < 2e-16 ***
itemi04  0.20627   0.04910  4.201 2.65e-05 ***
itemi05  2.41570   0.07970 30.311 < 2e-16 ***
itemi06  1.25343   0.05624 22.288 < 2e-16 ***
itemi07 -2.36083   0.07896 -29.899 < 2e-16 ***
itemi08  2.20807   0.07397 29.852 < 2e-16 ***
itemi09 -1.34192   0.05767 -23.267 < 2e-16 ***
itemi10  0.56369   0.05033 11.201 < 2e-16 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Correlation of Fixed Effects:
          item01 item02 item03 item04 item05 item06 item07 item08 item09
itemi02  0.107
itemi03  0.104  0.112
itemi04  0.109  0.117  0.114
itemi05  0.070  0.074  0.072  0.074
itemi06  0.097  0.104  0.102  0.105  0.068
itemi07  0.064  0.071  0.068  0.074  0.042  0.062
itemi08  0.075  0.079  0.078  0.080  0.054  0.073  0.045
itemi09  0.089  0.098  0.094  0.102  0.059  0.086  0.067  0.064
itemi10  0.107  0.115  0.112  0.118  0.074  0.104  0.071  0.080  0.098

> rasch_thetahat = coef(rasch)$person[, 1] # extracting respondent estimates
> head(rasch_thetahat,6)
[1]  0.56058142 -0.04968148 -0.04968148 -0.33875081  0.88702734  0.56058142

> # DIF analysis, rasch model no intercept
> raschDIF=glmer(x~0+race*item+(1|person),family=binomial("logit"),data=stacked,
  control=optmzrinfo)

> anova(rasch,raschDIF)
  Data: stacked
  Models:
rasch: x ~ 0 + item + (1 | person)

```

(continued)

TABLE 13.9. (*continued*)

```

raschDIF: x ~ 0 + race * item + (1 | person)
      Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
rasch    11 21298 21385 -10638     21276
raschDIF 21 20053 20219 -10005     20011 1265.4      10 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(raschDIF)
Generalized linear mixed model fit by maximum likelihood
  (Laplace Approximation) [glmerMod]
Family: binomial ( logit )
Formula: x ~ 0 + race * item + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC logLik deviance df.resid
20052.5 20218.8 -10005.3 20010.5     20279

Scaled residuals:
      Min      1Q Median      3Q      Max
-7.8289 -0.5727  0.2805  0.5285  5.6636

Random effects:
 Groups Name        Variance Std.Dev.
person (Intercept) 0.5452    0.7384

Number of obs: 20300, groups: person, 2030

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
race0       1.96990  0.09627 20.463 < 2e-16 ***
race1       0.45188  0.07011  6.445 1.16e-10 ***
itemi02    -0.49645  0.12278 -4.043 5.27e-05 ***
itemi03    -1.93684  0.11352 -17.062 < 2e-16 ***
itemi04    -1.09425  0.11645 -9.397 < 2e-16 ***
itemi05     1.75553  0.21515  8.160 3.36e-16 ***
itemi06    -0.25068  0.12676 -1.978  0.0480 *
itemi07    -4.58329  0.15299 -29.958 < 2e-16 ***
itemi08     1.38452  0.18827  7.354 1.93e-13 ***
itemi09    -3.14770  0.11972 -26.292 < 2e-16 ***
itemi10    -0.56587  0.12182 -4.645 3.40e-06 ***
race1:itemi02 -0.11563  0.15384 -0.752  0.4523
race1:itemi03  3.31383  0.15917 20.819 < 2e-16 ***
race1:itemi04  0.20540  0.14937  1.375  0.1691
race1:itemi05 -0.47056  0.24139 -1.949  0.0512 .
race1:itemi06  0.64497  0.15906  4.055 5.02e-05 ***
race1:itemi07  2.01717  0.19431 10.381 < 2e-16 ***
race1:itemi08 -0.28391  0.21589 -1.315  0.1885
race1:itemi09  1.19609  0.15927  7.510 5.92e-14 ***
race1:itemi10 -0.05070  0.15308 -0.331  0.7405
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation matrix not shown by default, as p = 20 > 12.
Use print(x, correlation=TRUE) or
  vcov(x)      if you need it

```

(continued)

TABLE 13.9. (continued)

> raschDIF_items = data.frame(easiness = fixef(raschDIF))	# extracting item estimates
> raschDIF_items	
	easiness
race	-1.51804306
itemi01	1.96992742
itemi02	1.47344736
itemi03	0.03306293
itemi04	0.87565008
itemi05	3.72542973
itemi06	1.71922697
itemi07	-2.61341883
itemi08	3.35441478
itemi09	-1.17781342
itemi10	1.40403013
race:itemi02	-0.11560624
race:itemi03	3.31385497
race:itemi04	0.20543369
race:itemi05	-0.47053486
race:itemi06	0.64499242
race:itemi07	2.01722988
race:itemi08	-0.28388044
race:itemi09	1.19613618
race:itemi10	-0.05067286

mix results fit information is very close to that of glmer, as is the person variability of 0.5877. The item easiness estimates ($\hat{\delta}_1^E = 1.11138$, $\hat{\delta}_2^E = 0.59220$, ..., $\hat{\delta}_{10}^E = 0.56369$) are correlated 0.99999 with those of glimmix's no DIF calibration with differences typically occurring in the fourth decimal place.

For our DIF model we introduce the race predictor information into our model through race*item while still treating item as fixed and person as random (i.e., raschDIF = glmer(x~0 + race*item + (1|person), . . .)). We compare our rasch and raschDIF fit results using the anova function. The DIF model-data fit θ is significantly better than the no DIF model: $\Delta G^2 = G_{\text{null}}^2 - G_{\text{DIF}}^2 = 21,275.9 - 20,010.5 = 1265.40$ with 10 dfs; lme4 labels G^2 as deviance and ΔG^2 as Chisq. Our BIC value also favors the DIF model over the no DIF model ($BIC_{\text{DIF}} = 20,219$ vs. $BIC_{\text{null}} = 21,385$). The introduction of examinee race information accounts for some of the person variability that leads to a reduction in person variability from 0.5877 to 0.5452.

The Fixed effects table provides our item and race effects. The effect of being a member of the Reference group is 1.96990, and that of being in the Focal group 0.45188 for a difference favoring the Reference group is 1.51802. Additionally and as was the case above, item 1 is substantially easier for the Reference group ($\hat{\delta}_{1,R}^E = \text{race0} = 1.9699$) than for the Focal group ($\hat{\delta}_{1,F}^E = \text{race1} = 0.45188$). We can obtain our Reference group's item location estimates from these Fixed effects table by taking into account our reference item (item 1). For example, $\hat{\delta}_{2,R}^E = 1.9669 + (-0.49645) = 1.47345$, $\hat{\delta}_{3,R}^E = 1.9669 + (-1.93684) = 0.03306$, and so on. For our Focal group our item location estimates

take into account not only the Focal group effect (`race1`) and the item's estimated location, but also the corresponding `race1:itemj` interaction (e.g., `race1:itemi02`). For instance, $\hat{\delta}_{2,F}^E = 0.45188 + (-0.49645) + (-0.11563) = -0.1602$, $\hat{\delta}_{3,F}^E = 0.45188 + (-1.93684) + 3.31383 = 1.82887$, and so on. These values are extremely close to those of glimmix (item*race Least Squares Means table) with a perfect correlation for the corresponding Reference and Focal groups; glmer: $\bar{\delta}_R^E = 1.0764$ ($SD = 1.8305$), $\bar{\delta}_F^E = 0.2040$ ($SD = 1.2743$) and glimmix: $\bar{\delta}_R^E = 1.0789$ ($SD = 1.8350$), $\bar{\delta}_F^E = 0.2043$ ($SD = 1.2747$). Alternatively, we can use the `fixed` function to extract the Reference group's item parameter estimates, $\hat{\delta}_{j,R}^E$ s. As is the case with Fixed effects, to obtain the $\hat{\delta}_{j,F}^E$ s requires that we take into account the race, item, and item-race interaction effects: $\hat{\delta}_{j,F}^E = (\text{item } j \text{ effect}) + (\text{race effect}) + (\text{item } j * \text{race effect})$. For example, $\hat{\delta}_{1,F}^E = 1.969927 + (-1.51804) + 0 = 0.45188$, $\hat{\delta}_{2,F}^E = 1.969927 + (-1.51804) + (-0.11561) = -0.1602$, etcetera. The *z* test (Estimate/Std. error) evaluates whether the item's estimated location or the item by group interaction is significantly different from zero.

The `race1:itemj` interaction can be thought of as an offset from the Reference group's corresponding item's estimated location that is due to the effect of race. Thus, for the Focal group items, the corresponding interaction terms' "z value" provides an indication of the presence of differential item function. If a Focal group item's z value is significant, then we have evidence that the item is performing differentially across racial groups. For our example, because the significant z value's for items 3, 6, 7, and 9 indicate these items are exhibiting DIF. Moreover, items 3, 6, 7 are easier for the Focal group than for the Reference group (e.g., $\hat{\delta}_{3,F}^E = 1.82887$ vs. $\hat{\delta}_{3,R}^E = 0.03306$); items 3 and 7 items are B items.

Person-Level Predictors for Respondents

Assume we calibrate a nutrition literacy scale and obtain respondent location estimates. The estimates' variability may be accounted for by a number of factors, such as number of nutrition courses taken, exercise regimen followed, and diet type. Accordingly, we regress our person location estimates on one or more predictor(s) to determine which predictors are comparatively more important. In this analytical approach as well as with other general linear model-based statistical analysis techniques, there is an assumption that the errors are independent and are identically distributed across cases. However, in Figure 4.3 we saw a quadratic relationship between SEE ($s_e(\hat{\theta})$) and person location estimates with smaller SEEs around $\theta = 0.4$ and progressively larger SEEs as we moved away from this point in both directions. This parabolic pattern indicates that our estimation errors are not identically distributed across our scale and, as a result, across our respondents. (The same issue arises with CTT in that the conditional standard errors of measurement are, generally speaking, nonconstant across the observed score scale [see Kolen et al. (1992)].) When we perform a regression analysis using person location estimates as the criterion, we tend to underestimate the effect(s) of our predictor(s).

(Adams et al., 1997; Mislevy, 1984). However, by using a multilevel approach, we can obtain more accurate estimates of our regression coefficient(s). Furthermore, a multilevel approach will yield estimates of our item and person locations as well as estimates of the strength of the relationship(s) between our person location estimates and our predictor(s) in a single step.

For person-level predictors (covariates) for respondents and following Kamata (2001), Equation 13.11 becomes

$$\beta_{0i} = \zeta_{00} + \sum_{s=i}^w \zeta_{0s} W_{si} + \theta_{0i}^{rW}, \quad (13.23)$$

where respondent i 's person location estimate, θ_{0i}^{rW} , is adjusted for the use of one or more person-level predictors ($W_{si}(s)$), and the $\zeta_{0s}(s)$ are the corresponding predictor effect(s); θ_{0i}^r does not involve person-level predictors and is thus unadjusted. Combining our level-1 and level-2 models, we obtain

$$\begin{aligned} p_{ji} &= \frac{\exp[\vartheta_{ji}]}{1 + \exp[\vartheta_{ji}]} = \frac{\exp[(\sum_{s=i}^w \zeta_{0s} W_{si} + \theta_{0i}^{rW}) - (-\zeta_{j0} - \zeta_{00})]}{1 + \exp[(\sum_{s=i}^w \zeta_{0s} W_{si} + \theta_{0i}^{rW}) - (-\zeta_{j0} - \zeta_{00})]} \\ &= \frac{\exp[\theta_{0i}^r - \delta_j]}{1 + \exp[\theta_{0i}^r - \delta_j]}. \end{aligned} \quad (13.24)$$

Recall that with the 1PL/Rasch model respondents with the same X receive the same location estimate. However, in the current formulation, respondents with the same X may differ in their location estimates θ_{0i}^{rW} 's, depending on whether they differ on their person-level predictor values. In some situations, this may be a reasonable outcome. For example, assume that we have test anxiety intensity as a person-level predictor (W) and two examinees receive the same observed score on a math test. However, one examinee (A) has higher test anxiety than the other examinee (B). Because the individual with the higher test anxiety obtained the same observed score as the individual with the lower test anxiety, the former might be considered to have a higher math ability than the latter. That is, A has higher ability than B because A 's higher test anxiety adversely affected A 's performance to yield the same observed score as B 's. If we take into account A 's higher test anxiety in our proficiency estimation, then A 's estimated location will be higher than that of B despite the two individuals obtaining the same observed score. From a mathematical perspective, the discrepancy between θ_{0i}^r and θ_{0i}^{rW} is a function solely of the respondents' predictor values. Thus, we have (Kamata, 2001)

$$\theta_{0i}^r - \theta_{0i}^{rW} = \sum_{s=i}^w \zeta_{0s} W_{si} \quad (\text{i.e., } \theta_{0i}^r = \sum_{s=i}^w \zeta_{0s} W_{si} + \theta_{0i}^{rW}).$$

Example: Person-Level Predictors for Respondents— Nutrition Literacy, proc glimmix

To demonstrate the use of person-oriented predictors for respondents, assume we are interested in nutrition literacy. Nutrition literacy is defined as “the degree to which individuals can obtain, process, and understand the basic health (nutrition) information and services they need to make appropriate health (nutrition) decisions” (Silk et al., 2008, p. 4). Zoellner et al. (2011) found relationships between participants’ diet quality and their nutrition literacy, age, gender, and participation in the Supplemental Nutrition Assistance Program (SNAP). Accordingly, we design a nutrition education program to help families make healthy food choices; such a program could be part of a SNAP-focused project of a Land Grant university’s Extension Division. We conceptualize nutrition literacy as a continuous latent variable. To assess the effectiveness of the program, we develop a nutrition literacy scale. We proceed through the steps of scale construction to develop a 10-item multiple-choice format instrument that measures participants’ nutrition literacy, with higher scores reflecting greater literacy than do lower scores. An example item from the scale could be the following.

Butter has lots of _____ fat that can increase cholesterol.

- (a) monounsaturated
- (b) polyunsaturated
- (c) saturated
- (d) trans

(This item is similar to one found on Diamond’s [2004] Nutrition Literacy Scale.)

We collect data from 1000 respondents who are primarily responsible for purchasing a household’s groceries and cooking meals. In addition to assessing nutrition literacy, we obtain information on our participants’ educational level. We operationalize educational level (`edlevel`) as the number of years of education except for postbaccalaureate degrees in which a master’s degree is coded as an 18 and a doctorate is coded as a 21 regardless of the number of years taken to obtain these degrees. The educational levels across our respondents ranged from 13 to 21 years with an average of almost 17 years.

Table 13.10 contains two `glimmix` programs; the data steps from Table 13.1 (not shown) are used to create the stacked _ data set from the file NutrLit.dat. The first program contains our person-level covariate, whereas the second program does not. On our first program’s `proc` statement, we specify maximum likelihood estimation approximation (`method = Laplace`) and include our respondents (`person`) and items (`item`) in our `class` statement.²² On our `model` statement, we specify our fixed effects covariate (`edlevel`) and request the fixed effects results by specifying `solution`; we use the `event = '1'` syntax in lieu of the equivalent `desc` option used above.

Tables 13.11 and 13.12 contain our corresponding output. The Fit Statistics tables present the deviance statistics for the models with the `edlevel` covariate (Table

TABLE 13.10. proc glimmix Programs with and without Person Covariate

```

proc glimmix data=stacked_data noclprint method=laplace;
  title "fixed items & random P w/ edlevel covariate";
  class item person;
  model x(event='1')= item edlevel / dist=binary link=logit solution;
  random intercept / subject=person solution;
  lsmeans item /ilink;
run;

proc glimmix data=stacked_data noclprint method=laplace;
  title "fixed items & random P, no covariate ";
  class item person;
  model x(event='1')= item / dist=binary link=logit solution;
  random intercept / subject=person solution;
  lsmeans item /ilink
run;

```

13.11, full model: $-2\ln L_F = G_F^2 = 11,146.08$) and without the covariate (Table 13.12, reduced model: $-2\ln L_R = G_R^2 = 11,158.59$). From Equation 6.7 we have

$$\Delta G^2 = (-2\ln L_R) - (-2\ln L_F) = G_R^2 - G_F^2 = 11,158.59 - 11,146.08 = 12.51$$

Our significant $\Delta G^2 (\chi_{(1),0.05}^2 = 3.84)$ indicates that using the covariate provides a significant improvement in model–data fit over not using the edlevel covariate. This interpretation is also supported by all the information criteria showing smaller values for the edlevel covariate model than for the no covariate model.

The effect of using the person covariate is evident when we compare the estimates of the variance of the random respondent with and without the covariate. Our variance estimate of 0.5463 without the edlevel covariate (Table 13.12: Covariance Parameter Estimates table) is reduced to 0.5330 when we introduce our edlevel covariate (Table 13.11: Covariance Parameter Estimates table). Moreover, from the Solution for Fixed Effects table we see that our person covariate estimate, $\hat{\zeta}_{01} = 0.1013$, is significantly different from 0 at the 5% significance level.

The Solution for Random Effects table presents our $\hat{\theta}_{0i}^{rw}$ s. That is, our first five respondents have nutrition literacy estimates of $-0.3612, -0.3612, -0.2610, -0.3612$, and 0.0079 . In contrast, the first four of these respondents are estimated to be located slightly higher when we do not take into account their education level ($\hat{\theta}_{01}^r = \hat{\theta}_{02}^r = \hat{\theta}_{03}^r = \hat{\theta}_{04}^r = -0.3128$; see Table 13.12). These four respondents have the same X score and thus receive the same $\hat{\theta}_{0i}^r$. In contrast, the $\hat{\theta}_{0i}^{rw}$ s for these respondents show two different estimates because of two different educational levels for respondent 3 and respondents 1, 2, and 4. The fifth respondent is estimated to have a higher nutrition literacy when we take into account their educational level than when do not utilize this

TABLE 13.11. proc glimmix Output with Person Covariate

Fit Statistics						
-2 Log Likelihood						11146.08
AIC (smaller is better)						11170.08
AICC (smaller is better)						11170.12
BIC (smaller is better)						11228.98
CAIC (smaller is better)						11240.98
HQIC (smaller is better)						11192.47
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error			
Intercept	person	0.5330	0.05555			
Solutions for Fixed Effects						
Effect	item	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		-3.7050	0.4970	8991	-7.46	<.0001
item	1	4.2376	0.1426	8991	29.71	<.0001
item	2	3.5842	0.1291	8991	27.77	<.0001
item	3	3.1698	0.1235	8991	25.66	<.0001
item	4	2.5118	0.1182	8991	21.25	<.0001
item	5	2.1516	0.1168	8991	18.42	<.0001
item	6	1.7668	0.1164	8991	15.18	<.0001
item	7	1.6476	0.1165	8991	14.15	<.0001
item	8	1.1869	0.1179	8991	10.06	<.0001
item	9	0.5242	0.1237	8991	4.24	<.0001
item	10	0
edlevel		0.1013	0.02860	8991	3.54	0.0004
Solution for Random Effects						
Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	person 1	-0.3612	0.5196	8991	-0.70	0.4870
Intercept	person 2	-0.3612	0.5196	8991	-0.70	0.4870
Intercept	person 3	-0.2610	0.5201	8991	-0.50	0.6158
Intercept	person 4	-0.3612	0.5196	8991	-0.70	0.4870
Intercept	person 5	0.007932	0.5185	8991	0.02	0.9878
item Least Squares Means						
item	Estimate	Standard Error	DF	t Value	Pr > t	Standard Error Mean
1	2.2509	0.1061	8991	21.21	<.0001	0.9047
2	1.5974	0.08767	8991	18.22	<.0001	0.8317
3	1.1830	0.07994	8991	14.80	<.0001	0.7655
4	0.5251	0.07278	8991	7.21	<.0001	0.6283
5	0.1649	0.07125	8991	2.31	0.0207	0.5411
6	-0.2200	0.07140	8991	-3.08	0.0021	0.4452
7	-0.3392	0.07182	8991	-4.72	<.0001	0.4160
8	-0.7999	0.07514	8991	-10.65	<.0001	0.3100
9	-1.4626	0.08502	8991	-17.20	<.0001	0.1881
10	-1.9868	0.09783	8991	-20.31	<.0001	0.1206

TABLE 13.12. proc glimmix Output without Person Covariate

Fit Statistics						
-2 Log Likelihood						11158.59
AIC (smaller is better)						11180.59
AICC (smaller is better)						11180.62
BIC (smaller is better)						11234.58
CAIC (smaller is better)						11245.58
HQIC (smaller is better)						11201.11
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error			
Intercept	person	0.5463	0.05628			
Solutions for Fixed Effects						
Effect	item	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		-1.9860	0.09780	8991	-20.31	<.0001
item	1	4.2377	0.1426	8991	29.71	<.0001
item	2	3.5835	0.1290	8991	27.78	<.0001
item	3	3.1685	0.1235	8991	25.66	<.0001
item	4	2.5099	0.1181	8991	21.25	<.0001
item	5	2.1496	0.1167	8991	18.42	<.0001
item	6	1.7647	0.1163	8991	15.18	<.0001
item	7	1.6455	0.1164	8991	14.14	<.0001
item	8	1.1851	0.1178	8991	10.06	<.0001
item	9	0.5233	0.1236	8991	4.23	<.0001
item	10	0
Solution for Random Effects						
Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	person 1	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 2	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 3	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 4	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 5	-0.04072	0.5215	8991	-0.08	0.9378
item Least Squares Means						
item	Estimate	Standard Error	DF	t Value	Pr > t	Standard Error Mean
1	2.2517	0.1062	8991	21.19	<.0001	0.9048
2	1.5975	0.08780	8991	18.19	<.0001	0.8317
3	1.1825	0.08005	8991	14.77	<.0001	0.7654
4	0.5239	0.07288	8991	7.19	<.0001	0.6281
5	0.1636	0.07134	8991	2.29	0.0218	0.5408
6	-0.2213	0.07149	8991	-3.10	0.0020	0.4449
7	-0.3405	0.07190	8991	-4.74	<.0001	0.4157
8	-0.8009	0.07521	8991	-10.65	<.0001	0.3098
9	-1.4627	0.08504	8991	-17.20	<.0001	0.1880
10	-1.9860	0.09780	8991	-20.31	<.0001	0.1207
						0.01038

information (i.e., full model: $\hat{\theta}_{05}^{rW} = 0.0079$, reduced model: $\hat{\theta}_{05}^r = -0.0407$). Nevertheless, when we predict $\hat{\theta}_{0i}^r$ from our educational level covariate model (i.e., $\zeta_{01}(edlevel) + \hat{\theta}_{0i}^{rW}$) and correlate it with the no predictor model's $\hat{\theta}_{0i}^r$'s the two sets of estimates show strong agreement ($r = 0.998$). Figure 13.3 shows the corresponding scatterplot.

Although the respondents' different educational levels affect our respondents' nutrition literacy estimates, they do not affect the item parameter estimates. A comparison of the covariate model's item parameter estimates with those of the no covariate model's estimates show near-perfect agreement (Item Least Squares Means table). For example, for our covariate model, we have $\hat{\delta}_1^E = 2.2509$, $\hat{\delta}_2^E = 1.5974$, $\hat{\delta}_3^E = 1.1830$, etc. and for our no covariate model the estimates are $\hat{\delta}_1^E = 2.2517$, $\hat{\delta}_2^E = 1.5975$, $\hat{\delta}_3^E = 1.1825$, etc.

Example: Person-Level Predictors for Respondents, lme4

Table 13.13 shows our R session. After reading and transforming our data, we perform a standard Rasch analysis (i.e., the no covariate model), `rasch = glmer(x~0 + item + (1|person), ...)`. Our information criteria are an AIC of 11,180.6 and a BIC value of 11,259.9 with a G_R^2 of 11,158.6. The person variability is 0.5457 and item location estimates of $\hat{\delta}_1^E = 2.25156$, $\hat{\delta}_2^E = 1.597285$, $\hat{\delta}_3^E = 1.18227$, and so on, are virtually identical to those from `glimmix` (any differences between the two sets appear in the fourth

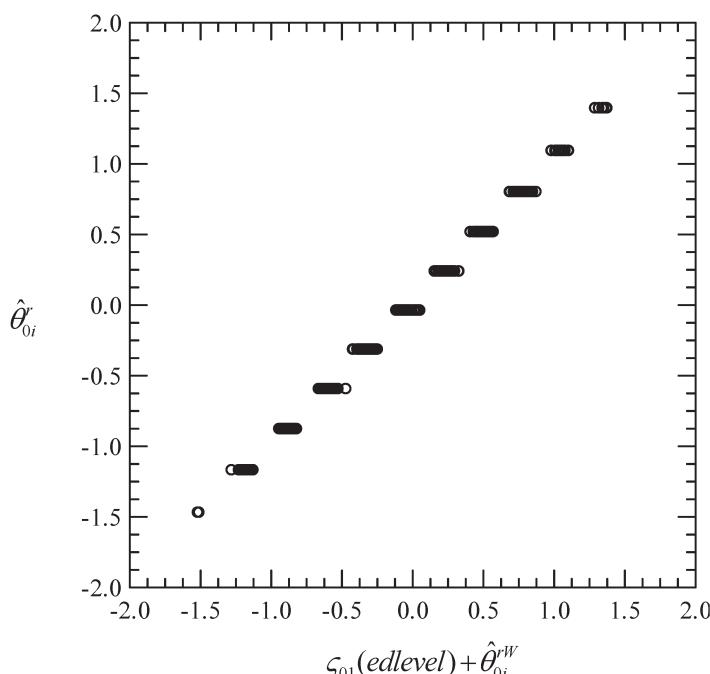


FIGURE 13.3. Scatterplot of $\zeta_{01}(edlevel) + \hat{\theta}_{0i}^{rW}$ and predicted $\hat{\theta}_{0i}^r$.

TABLE 13.13. lme4 Session for Person Covariate Analysis

```

> library(tidyverse); library(lme4); library(optimx); library(dplyr)

> unstacked = read.table(file.choose(), header=F) # read data: Nutrlit.dat

> # create meaningful variable names
> names(unstacked)=c('person','PwiEdctr','educator','edlevel','i01','i02','i03',
+                     'i04', 'i05', 'i06','i07','i08','i09','i10')

> head(unstacked,10)
  person PwiEdctr educator edlevel i01 i02 i03 i04 i05 i06 i07 i08 i09 i10
1      1         1        1     18    1   1   0   1   0   0   1   0   0   0   0
2      2         2        2     18    1   1   0   0   1   1   0   0   0   0   0
:
10     10        10       10     18    1   0   1   0   0   0   1   0   0   0   0

> # For this example we don't need PwiEdctr & educator. To remove these variables
> # we use 'select' function to identify the variables we want to keep.
> # Alternatively, we can use 'rm' and 'within' to remove a variable or use
> # indexing. For example, to remove only PwiEdctr: 'within(unstacked,
> # rm(PwiEdctr))' or 'unstacked=unstacked[,-2]'

> unstacked=select(unstacked, person, edlevel, i01, i02, i03, i04,
+                   i05, i06, i07, i08, i09, i10) # select is from dplyr package

> head(unstacked,10)
  person edlevel i01 i02 i03 i04 i05 i06 i07 i08 i09 i10
1      1     18    1   1   0   1   0   0   1   0   0   0
2      2     18    1   1   0   0   1   1   0   0   0   0
:
10     10    18    1   0   1   0   0   0   1   0   0   0

> # reformat unstacked data into stacked format & reorder data to be items
w/i person
> tmpstacked = gather(unstacked,key=item,value=x,i01:i10)
> stacked = arrange(tmpstacked,person,edlevel,item)

># to speed up execution optmzrinfo - see Table 13.5

> # doing the rasch model w/o intercept
> rasch=glmer(x~0+item+(1|person),family=binomial("logit"),data=stacked, control=
+ optmzrinfo)

> summary(rasch)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation)
[glmerMod]
Family: binomial ( logit )
Formula: x ~ 0 + item + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC      logLik deviance df.resid
11180.6  11259.9  -5579.3  11158.6      9989

Scaled residuals:
    Min      1Q  Median      3Q      Max
-4.5615 -0.6676  0.2520  0.6544  4.1605

```

(continued)

TABLE 13.13. (*continued*)

```

Random effects:
 Groups Name           Variance Std.Dev.
 person (Intercept) 0.5457   0.7387
 Number of obs: 10000, groups: person, 1000

Fixed effects:
    Estimate Std. Error z value Pr(>|z|)
itemi01  2.25156   0.10749 20.948 < 2e-16 ***
itemi02  1.59728   0.08718 18.322 < 2e-16 ***
itemi03  1.18227   0.07874 15.015 < 2e-16 ***
itemi04  0.52378   0.07105  7.372 1.68e-13 ***
itemi05  0.16357   0.06943  2.356  0.01848 *
itemi06 -0.22127   0.06960 -3.179  0.00148 **
itemi07 -0.34040   0.07004 -4.860 1.17e-06 ***
itemi08 -0.80080   0.07357 -10.885 < 2e-16 ***
itemi09 -1.46260   0.08419 -17.372 < 2e-16 ***
itemi10 -1.98590   0.09818 -20.227 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      item01 item02 item03 item04 item05 item06 item07 item08 item09
itemi02  0.062
itemi03  0.068  0.083
itemi04  0.073  0.090  0.099
itemi05  0.074  0.091  0.100  0.111
itemi06  0.072  0.089  0.099  0.110  0.113
itemi07  0.071  0.088  0.098  0.109  0.112  0.112
itemi08  0.067  0.083  0.092  0.103  0.107  0.107  0.107
itemi09  0.057  0.071  0.079  0.090  0.093  0.094  0.094  0.090
itemi10  0.048  0.060  0.067  0.076  0.079  0.080  0.080  0.078  0.070

> rasch_thetahat = coef(rasch)$person[, 1]                                     # extracting respondent est

> head(rasch_thetahat,50)
[1] -0.31261922 -0.31261922 -0.31261922 -0.31261922 -0.04068828
[6] -0.58678953 -0.04068828 -0.04068828 -0.86536661 -0.58678953
:
[46] -0.31261922 -0.86536661 -0.58678953 -0.31261922 -0.31261922

> # introducing edlevel person predictor
> raschPersonEdlev =glmer(x~0+item+edlevel+(1|person),family=binomial("logit"),
  data=stacked, control=optmzrinfo)

> anova(rasch,raschPersonEdlev)
Data: stacked
Models:
rasch: x ~ 0 + item + (1 | person)
raschPersonEdlev: x ~ 0 + item + edlevel + (1 | person)
              Df AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
rasch          11 11181 11260 -5579.3     11159
raschPersonEdlev 12 11170 11257 -5573.0     11146 12.51      1  0.0004048 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(continued)

TABLE 13.13. (*continued*)

```

> summary(raschPersonEdlev)
Generalized linear mixed model fit by maximum likelihood
  (Laplace Approximation)
[glmerMod]
Family: binomial ( logit )
Formula: x ~ 0 + item + edlevel + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC   logLik deviance df.resid
11170.1 11256.6 -5573.0  11146.1     9988

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.6625 -0.6710  0.2462  0.6623  3.9245

Random effects:
Groups Name        Variance Std.Dev.
person (Intercept) 0.5326   0.7298
Number of obs: 10000, groups: person, 1000

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
itemi01  0.53322   0.49190  1.084 0.278363
itemi02 -0.12020   0.48791 -0.246 0.805409
itemi03 -0.53469   0.48655 -1.099 0.271793
itemi04 -1.19260   0.48563 -2.456 0.014058 *
itemi05 -1.55266   0.48564 -3.197 0.001388 **
itemi06 -1.93750   0.48599 -3.987 6.70e-05 ***
itemi07 -2.05666   0.48618 -4.230 2.33e-05 ***
itemi08 -2.51733   0.48720 -5.167 2.38e-07 ***
itemi09 -3.17998   0.48968 -6.494 8.36e-11 ***
itemi10 -3.70420   0.49281 -7.516 5.63e-14 ***
edlevel  0.10124   0.02837  3.569 0.000358 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          item01 item02 item03 item04 item05 item06 item07 item08 item09 item10
itemi02  0.963
itemi03  0.965  0.973
itemi04  0.968  0.976  0.979
itemi05  0.968  0.976  0.979  0.981
itemi06  0.968  0.976  0.979  0.981  0.982
itemi07  0.968  0.976  0.979  0.981  0.982  0.982
itemi08  0.967  0.975  0.978  0.980  0.981  0.981  0.981
itemi09  0.963  0.971  0.974  0.977  0.977  0.977  0.977  0.976
itemi10  0.958  0.966  0.969  0.972  0.972  0.972  0.972  0.971  0.968
edlevel -0.976 -0.984 -0.987 -0.989 -0.990 -0.990 -0.990 -0.989 -0.985 -0.980

> raschEdlev_items = data.frame(easiness = fixef
  (raschPersonEdlev))                                     # extracting item estimates

```

(continued)

TABLE 13.13. (continued)

```

> raschEdlev_items
      easiness
itemi01  0.5332192
itemi02 -0.1201988
itemi03 -0.5346949
itemi04 -1.1926043
itemi05 -1.5526610
itemi06 -1.9374980
itemi07 -2.0566555
itemi08 -2.5173306
itemi09 -3.1799809
itemi10 -3.7042020
edlevel  0.1012397

> # extracting person estimates
> raschEdlev_thetahat = coef(raschPersonEdlev)$person[, 1]
> peoplehEdlev=matrix(raschEdlev_thetahat)  # reformatting person estimates

> head(peoplehEdlev,5)
      [,1]
[1,] -0.361003198
[2,] -0.361003198
[3,] -0.260919090
[4,] -0.361003198
[5,]  0.007914445

> tail(peoplehEdlev,2)
      [,1]
[999,]  0.226222007
[1000,] -0.532248646

```

decimal place). Our person location estimates (`rasch_thetahat`) show the first four persons located at -0.31262 and the fifth at -0.0407 .

Introducing our `edlevel` person-level predictor (i.e., `raschPersonEdlev = glmer(x~0 + item + edlevel + (1|person), ...)`) leads to better model–data fit than with the Rasch model. That is, we see a reduction in AIC and BIC to 11,170.1 and 11,256.6, respectively, with a $G_F^2 = 11,146.1$. Comparing the two models (`anova(rasch, raschPersonEdlev)`) shows the $\Delta G^2 = 12.51$ is significant at the 5% significance level. The estimated person variability is reduced to 0.53264 (cf. `glimmix`'s estimated variance is 0.5330). The item location estimates are $\hat{\delta}_1^E = 0.53322$, $\hat{\delta}_2^E = -0.12020$, $\hat{\delta}_3^E = -0.53469$, etc., with a significant covariate (ζ_{01}) effect estimated to be 0.10124 (cf. `glimmix`'s $\zeta_{01} = 0.1013$). Unlike the close correspondence observed between the `glimmix` no covariate (Rasch) and the person covariate models' item location estimates (e.g., Rasch: $\hat{\delta}_1^E = 2.2517$, $\hat{\delta}_2^E = 1.5975$; covariate model: $\hat{\delta}_1^E = 2.2509$, $\hat{\delta}_2^E = 1.5974$), these `glmer` item estimates appear to be vastly different from the no covariate (Rasch) analysis. However, the no covariate and covariate models' estimates are on different scales. The two estimate sets have a correlation of 0.9999998. When we transform our covariate model's estimates to be on the same scale as the no covariate model, our estimates are

highly similar (e.g., covariate model: $\hat{\delta}_1^E = 2.25035$, $\hat{\delta}_2^E = 1.59690$, $\hat{\delta}_3^E = 1.18238$, and so forth).

Our nutrition literacy person location estimates are consistent with those of glimmix. For instance, the $\hat{\theta}_{0i}^{rW}$'s for the first five respondents have nutrition literacy estimates of -0.3610 , -0.3610 , -0.2609 , -0.3610 , and 0.0079 . In contrast, the first four of these respondents are estimated to be located slightly higher when we do not take into account their education level ($\hat{\theta}_{01}^r = \hat{\theta}_{02}^r = \hat{\theta}_{03}^r = \hat{\theta}_{04}^r = -0.3126$). As stated above, the $\hat{\theta}_{0i}^{rW}$'s for these respondents show two estimates because there are two different educational levels across these four respondents. The fifth respondent is estimated to have a higher nutrition literacy when we take into account their educational level than when we do not utilize this information (i.e., covariate model: $\hat{\theta}_{05}^{rW} = 0.0079$, no covariate model: $\hat{\theta}_{05}^r = -0.0407$). The person nutritional literacy estimates show a perfect correlation with those of glimmix.

Item-Level Predictors for Items

Above we examined our vocabulary test items for the presence of DIF. In doing so, we utilized a person characteristic to account for the possibility that an item might differ in locations with respect to the person characteristic (e.g., race). We now discuss accounting for item variability in terms of item characteristics. One approach mentioned above was the LLTM (Fischer, 1973; see Appendix E). In the LLTM, we use item characteristics as predictors of item location parameters. As such, the incorporation of item-level predictors into IRT models is not new. Below we use the LLTM for presenting the ideas underlying the use of item-level predictors.

The LLTM is an extension of the Rasch model designed to look at the impact of item characteristics on item locations. These item characteristics can include cognitive operations/skills, item features, item response format, instructional conditions, item position, and so on (see Embretson, 1984; Kubinger, 2009).²³ In the LLTM, the Rasch model's item location parameter is a linear function of a common set of basic parameters (e.g., cognitive operations) that perfectly describe performance on the item (see Van den Noortgate, De Boeck, & Meuldres, 2003).

With the LLTM item j 's location is a weighted linear composite of item characteristics (e.g., elementary components) that describe performance on item j that after centering the weighted components is

$$\delta_j = \sum_{s=1}^S q_{js} \eta_s, \quad (13.25)$$

where η_s is a basic parameter associated with elementary component s (e.g., a cognitive operation required to answer an item), S is the number of components, and q_{js} is a weight for component s for item j ; the q_{js} 's may or may not be indicator variables for the item characteristics. Incorporating Equation 13.25 into our Rasch model and after centering our composite yields the LLTM

$$p_{ji} = \frac{\exp[\theta_i - \delta_j]}{1 + \exp[\theta_i - \delta_j]} = \frac{\exp[\theta_i - \sum_{s=1}^S q_{js} \eta_s]}{1 + \exp[\theta_i - \sum_{s=1}^S q_{js} \eta_s]}. \quad (13.26)$$

(See Appendix E for more detail on the LLTM.)

From a multilevel perspective, treating items as fixed and starting with Equation 13.7, we have for person i

$$\vartheta_{ij} = \beta_{0i} + \sum_{q=1}^{L-1} \beta_{qi} X_{qij}. \quad (13.27)$$

Because

$$\beta_{0i} = \zeta_{00} + \theta_{0i}^r, \quad (13.28)$$

and letting

$$\beta_{qi} = \sum_{s=1}^S q_{js} \eta_s, \quad (13.29)$$

then by substitution and Equation 13.15, the log odds of a response of 1 for person i on item j is

$$\vartheta_{ji} = \theta_{0i}^r - \sum_{s=1}^S q_{js} \eta_s \quad (13.30)$$

According to Equation 13.25, the item location is assumed to be perfectly predictable by the weighted item characteristics (i.e., $\hat{\delta}'_j = \sum q_{js} \hat{\eta}_s$). However, typically one finds that LLTM item location estimates ($\hat{\delta}'_j$), although well correlated with those of the Rasch model ($\hat{\delta}_j$), will not be the same as the $\hat{\delta}'_j$'s (i.e., $\hat{\delta}_j \neq \hat{\delta}'_j$). However, by adding an error term to Equation 25, we can capture the idiosyncratic item variability not accounted for by our item characteristics. Thus, rather than treat items as fixed effects, we allow them to vary and consider them random effects in our model. From this perspective, the items' locations are a function of the identified item characteristics plus some item-specific unique characteristic. The random effect captures the unique item variability associated with the discrepancy from the fixed effect. Janssen and De Boeck (as cited in Rijmen & De Boeck, 2002; also see Janssen, Schepers, & Peres, 2004) present such a model. In this random item effects model (LLTM+ ε ; De Boeck, 2008), our item locations are a function of item characteristics plus a unique item j characteristic (ε_j)

$$\delta_j = \sum_{s=1}^S q_{js} \eta_s + \varepsilon_j. \quad (13.31)$$

Because items are now treated as random, we have an associated distributional assumption for ε_j . Specifically, ε_j is assumed to be distributed as $N(0, \sigma_\varepsilon^2)$ where σ_ε^2

is the residual variance (Janssen et al., 2004) and is assumed constant across items. Therefore, we have

$$\vartheta_{ji} = \theta_{0i}^r - \left(\sum_{s=1}^S q_{js} \eta_s + \varepsilon_j \right). \quad (13.32)$$

Our randomly sampled items' locations are assumed to be distributed normally with mean δ'_j and common variance σ_ε^2 across items (i.e., $\delta_j \sim N(\delta'_j, \sigma_\varepsilon^2)$) (Janssen et al., 2004). Items with the same pattern of q_{js} (s) belong to the same item population ("item family," De Boeck, 2008) and are considered exchangeable (Janssen et al., 2004).

Although the introduction of the LLTM+ ε was motivated by the discrepancy between $\hat{\delta}_j$ and δ'_j , the concept of considering items as random is not new (cf. domain sampling, Tryon, 1935, 1957, 1959; Generalizability Theory, Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Additionally, one might consider that items created under an Automated Item Generation (see Embretson & Kingston, 2018; Gierl & Haladyna, 2013) framework to be sampled (i.e., created) from a domain or universe of potential items.²⁴ De Boeck (2008) provides a brief history and rationale for why items can sometimes be considered as random.

Example: Item-Level Predictors for Items— Nutrition Literacy, proc glimmix

Recall in our nutrition literacy example above that we developed a 10-item multiple-choice format instrument. These items are designed to involve two components. The first component reflects whether the item included technical terminology (e.g., terms such as trans fat, saturated fat, polyunsaturated fat), whereas the second component has to do with food safety (e.g., temperatures, food storage, food handling). Whether an item reflects one or both of these components is indicated by using a design or weight matrix. Our weight matrix (**Q**) is L x S

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

where the entries in **Q** reflect whether the item involves the component ("1") or not ("0"). For instance, items 1 and 2 include just technical terminology (i.e., $q_{11} = q_{12} = 1$, $q_{12} = q_{22} = 0$), items 3–5 include both technical terminology and food safety (e.g., $q_{31} = 1$,

$q_{32} = 1$), and item 6 reflects just food safety (i.e., $q_{61} = 0, q_{62} = 1$), and so on. Our data are stacked with one **Q** per respondent (see Figure 13.4).

Table 13.14 contains two **glimmix** programs; the data steps from Table 13.1 (not shown) are used to create the stacked _ data set from the file LLTM.dat. The first program specifies the LLTM+ ε model, whereas the second program is a no item covariate model (simple “Rasch”) treating both items and respondents as random. For our LLTM+ ε model, we specify our two components $q_{1,1}$ and $q_{1,2}$, a no intercept model (**noint**), and output the fixed effects solution for our components (**solution**) on our **model** statement. (For the LLTM analysis of these data, see Appendix E.) Unlike the **glimmix** programs above, we now have two random statements, with one for items (**item**) and the other for respondents (**person**). Thus, we are combining random item variation with random person variation (i.e., crossed random effects). (Because quadrature approximation cannot be used with crossed random effects models, we specify the Laplace approximation for maximum likelihood estimation on the **proc** statement.) The output for our analysis is shown in Table 13.15, with the output for the no item covariate found in Table 13.16.

Comparing the corresponding information indices (Fit Statistics tables), we see that the LLTM+ ε provides relatively better model–data fit than the model without the item components. Additionally, the difference in our deviance statistics shows that

Multilevel (stacked) Format

Person	item	x	q1	q2
1	1	1	1	0
1	2	1	1	0
1	3	0	1	1
1	4	1	1	1
1	5	0	1	1
1	6	0	0	1
1	7	0	0	1
1	8	0	1	1
1	9	0	0	1
1	10	0	0	1
2	1	1	1	0
2	2	1	1	0
2	3	0	1	1
2	4	0	1	1
2	5	1	1	1
2	6	1	0	1
2	7	1	0	1
2	8	1	1	1
2	9	1	0	1
2	10	1	0	1
:				

FIGURE 13.4. Schematic of stacked data input file for the first two respondents (L = 10).

TABLE 13.14. proc glimmix Program with Item Covariate (LLTM+ ε)

```

proc glimmix data=stacked_data noclprint method=laplace;
  title "random items & random P, LLTM + e; 2 components";
  class item person;
  model x(desc)= q1 q2 /noint dist=binary link=logit solution;
    random intercept / subject=item solution;
    random intercept / subject=person solution;
  run;

proc glimmix data=stacked_data noclprint  method=laplace;
  title "fixed items & random P Rasch ";
  class item person;
  model x(desc)= /noint  dist=binary link=logit solution;
    random intercept /subject=item solution;
    random intercept/subject=person solution;
  run;

```

the LLTM+ ε fits significantly better than our no item covariate model ($\Delta G^2 = 11.95$, $df = 2$).

Our $\hat{\eta}_s$ s are $\hat{\eta}_1 = 1.5906$ and $\hat{\eta}_2 = -1.1620$ (Solutions for Fixed Effects table). Both of our components are significant at the 5% significance level. Thus, the technical terminology and food safety predictors appear to be useful in accounting for the variability in our item location estimates. In other words, the difference in item 8's location ($\hat{\delta}_8^E = -1.1707$) and item 6's easiness ($\hat{\delta}_6^E = 0.9453$) is accounted for, in part, by their respective component profiles; see the Solution for Random Effects table ($\hat{\delta}_1^E = 0.5410$, $\hat{\delta}_2^E = -0.07903$, ..., $\hat{\delta}_{10}^E = -0.7120$). Moreover, the introduction of our item covariates reduces our item variance from 1.6007 (Table 13.16: Covariance Parameter Estimates table) to 0.4810 (Table 13.15: Covariance Parameter Estimates table). Our random respondent variance estimate with and without the item covariates is essentially 0.529.

Example: Item-Level Predictors for Items—Nutrition Literacy, lme4

We now reanalyze our nutrition literacy data using lme4. Because our data consists of values that are not delimited (e.g., “11101001000”), we use the `read.fortran` function to read our data. To read these fixed-format data, we pass the pseudo-FORTRAN format (`c("1I8","10I1")`) to tell the function to read one integer value that occupies eight columns (“1I8”), followed by 10 integer values each one column wide (“10I1”); for more information on FORTRAN Formats, see Appendix G, “FORTRAN Formats.” Following the reading of our data into the unstacked data frame, we create our Q matrix (see Appendix E), stack our data, and define our optimizer (see Table 13.13).

Table 13.17 presents our session. Our first analysis is a Rasch calibration of the data (i.e., `rasch = glmer(x~0 + (1|item) + (1|person), ...)`) in which we

TABLE 13.15. proc glimmix Output with Item Covariate (LLTM + ε)

Fit Statistics						
-2 Log Likelihood					11217.94	
AIC (smaller is better)					11225.94	
AICC (smaller is better)					11225.94	
BIC (smaller is better)					11217.94	
CAIC (smaller is better)					11221.94	
HQIC (smaller is better)					11217.94	
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error			
Intercept	item	0.4810	0.2179			
Intercept	person	0.5295	0.05498			
Solutions for Fixed Effects						
Effect	Estimate	Standard Error	DF	t Value	Pr > t	
q1	1.5906	0.3498	8991	4.55	<.0001	
q2	-1.1620	0.3030	8991	-3.84	0.0001	
Solution for Random Effects						
Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	item 1	0.5410	0.3565	8991	1.52	0.1291
Intercept	item 2	-0.07903	0.3542	8991	-0.22	0.8234
Intercept	item 3	0.6776	0.3077	8991	2.20	0.0277
Intercept	item 4	0.06206	0.3066	8991	0.20	0.8396
Intercept	item 5	-0.2733	0.3065	8991	-0.89	0.3725
Intercept	item 6	0.9453	0.3068	8991	3.08	0.0021
Intercept	item 7	0.8344	0.3069	8991	2.72	0.0066
Intercept	item 8	-1.1707	0.3071	8991	-3.81	0.0001
Intercept	item 9	-0.2180	0.3086	8991	-0.71	0.4801
Intercept	item 10	-0.7120	0.3108	8991	-2.29	0.0220
Intercept	person 1	-0.3027	0.5154	8991	-0.59	0.5570
Intercept	person 2	-0.3027	0.5154	8991	-0.59	0.5570
Intercept	person 3	-0.3027	0.5154	8991	-0.59	0.5570
Intercept	person 4	-0.3027	0.5154	8991	-0.59	0.5570
Intercept	person 5	-0.03832	0.5141	8991	-0.07	0.9406

:					
Intercept	person 999	0.2260	0.5153	8991	0.44 0.6609
Intercept	person 1000	-0.5693	0.5192	8991	-1.10 0.2729

TABLE 13.16. proc glimmix Output without Item Covariate ("Rasch" Model with Random Items and Random Respondents)

Fit Statistics						
						:
-2 Log Likelihood						11229.89
AIC (smaller is better)						11233.89
AICC (smaller is better)						11233.89
BIC (smaller is better)						11229.89
CAIC (smaller is better)						11231.89
HQIC (smaller is better)						11229.89
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error			
Intercept	item	1.6007	0.7200			
Intercept	person	0.5294	0.05497			
Solution for Random Effects						
Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	item 1	2.1299	0.1034	8991	20.61	<.0001
Intercept	item 2	1.5045	0.08517	8991	17.67	<.0001
Intercept	item 3	1.1105	0.07756	8991	14.32	<.0001
Intercept	item 4	0.4902	0.07057	8991	6.95	<.0001
Intercept	item 5	0.1528	0.06910	8991	2.21	0.0270
Intercept	item 6	-0.2073	0.06924	8991	-2.99	0.0028
Intercept	item 7	-0.3189	0.06964	8991	-4.58	<.0001
Intercept	item 8	-0.7511	0.07284	8991	-10.31	<.0001
Intercept	item 9	-1.3770	0.08245	8991	-16.70	<.0001
Intercept	item 10	-1.8761	0.09501	8991	-19.75	<.0001
Intercept	person 1	-0.3031	0.5153	8991	-0.59	0.5565
Intercept	person 2	-0.3031	0.5153	8991	-0.59	0.5565
Intercept	person 3	-0.3031	0.5153	8991	-0.59	0.5565
Intercept	person 4	-0.3031	0.5153	8991	-0.59	0.5565
Intercept	person 5	-0.03874	0.5140	8991	-0.08	0.9399
:						
Intercept	person 999	0.2256	0.5152	8991	0.44	0.6615
Intercept	person 1000	-0.5695	0.5191	8991	-1.10	0.2726

treat our items as random by specifying a varying intercept for item (i.e., $(1|item)$).²⁵ Our information criteria are an AIC of 11,233.9 and a BIC value of 11,248.3 with a G_R^2 of 11,229.9. The person variability is 0.5294 with estimated item variance of 1.6007 (see Random effects table). The location estimates (`itemest _`) of $\hat{\delta}_1^E = 2.1298815$, $\hat{\delta}_2^E = 1.5045406$, $\hat{\delta}_3^E = 1.1104772$, etc., are almost identical to the estimates obtained from `glimmix` (Table 13.16). Similarly, our person location estimates (`raschRnd _ thetahat`) agree with those of `glimmix` with the first four estimated locations of -0.30306342 , the fifth at -0.03873654 , and so on.

In our LLTM+ ε analysis, we treat our items as random but the components as fixed (`LLTMe = glmer(x~0 + (1|item) + q1 + q2 + (1|person), ...)`).²⁶ The associated fit information shows a slight decrease in AIC but an increase in BIC. Nevertheless, the $G_{LLTM+\varepsilon}^2$ of 11,217.9 is significantly less than that of the random item Rasch calibration with a ΔG^2 of 11.951 ($p = 0.002541$); see `anova(raschRnd, raschLLTMe)` results. Consequently, there is better model–data fit with the LLTM+ ε model than with the random item Rasch model.

By introducing the item covariates, our estimated item variance decreases to 0.4810 from the 1.6007 seen with the Rasch calibration. Our $\hat{\eta}_s$ s are found in the Fixed Effects table ($\hat{\eta}_1 = 1.5906$ and $\hat{\eta}_2 = -1.1620$) with location estimates (`itemLLTMe`) of $\hat{\delta}_1^E = 0.5410376$, $\hat{\delta}_2^E = -0.07900892$, $\hat{\delta}_3^E = 0.67764644$, and so on. Both the location estimates and the $\hat{\eta}_s$ s are identical or almost so to the values obtained from `glimmix` (Table 13.15). This is also the case for the respondent's estimated locations (`peopleLLTMe`).

Above we treated item-level predictors for items in separate models than those for person-level predictors for respondents and item. However, each of these can be considered a reduced version of a general model that has both item-level predictors for items and person-level predictors for respondents. For more information on such a general model, see De Boeck and Wilson (2004).

Multilevel IRT—Three Levels

As mentioned above, we can generalize our two-level model to a three-level model. This third level would have our respondents nested within some grouping variable, such as schools, classrooms, countries, and so forth. For example, assume that our third level represents the groups to which respondents belong. Following Kamata (2001), we have for our level-1 model logit for the j th item and the i th person who is nested within the g th group

$$\vartheta_{ijg} = \beta_{0ig} + \beta_{1ig} X_{1ijg} + \beta_{2ig} X_{2ijg} + \dots + \beta_{(L-1)ig} X_{(L-1)ijg}, \quad (13.33)$$

where our third subscript, g , reflects our group ($g = 1 \dots G$) and all other terms are as defined above. Our person-level (i.e., level-2) model is analogous to Equations 13.10 and

TABLE 13.17. lme4 Session for Calibration with Item Covariate (LLTM+ ε)

```

> library(tidyverse); library(lme4); library(optimx)

> unstacked = read.fortran("LLTM.dat",c("1I8","1O1I"))

:
# steps to create stacked file

> head(stacked,20)
  caseID person item x comp1 q1 comp2 q2
  1       1     1 i01 1 q101 1 q201 0
  2       2     1 i02 1 q102 1 q202 0
  3       3     1 i03 0 q103 1 q203 1
  4       4     1 i04 1 q104 1 q204 1
  5       5     1 i05 0 q105 1 q205 1
  6       6     1 i06 0 q106 0 q206 1
  7       7     1 i07 1 q107 0 q207 1
  8       8     1 i08 0 q108 1 q208 1
  9       9     1 i09 0 q109 0 q209 1
 10      10    1 i10 0 q110 0 q210 1
 11      11    2 i01 1 q101 1 q201 0
 12      12    2 i02 1 q102 1 q202 0
 13      13    2 i03 0 q103 1 q203 1
 14      14    2 i04 0 q104 1 q204 1
 15      15    2 i05 1 q105 1 q205 1
 16      16    2 i06 1 q106 0 q206 1
 17      17    2 i07 0 q107 0 q207 1
 18      18    2 i08 0 q108 1 q208 1
 19      19    2 i09 0 q109 0 q209 1
 20      20    2 i10 0 q110 0 q210 1

> # Rasch analysis (random items)
> raschRnd=glmer(x~0+(1|item)+(1|person),family=binomial("logit"),data=lltmMMdat,
  control=optmzrinfo)
> summary(raschRnd)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + (1 | item) + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC  logLik deviance df.resid
 11233.9  11248.3 -5614.9   11229.9      9998

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2480 -0.6781  0.2696  0.6678  3.8892

Random effects:
 Groups Name        Variance Std.Dev.
 person (Intercept) 0.5294  0.7276
 item   (Intercept) 1.6007  1.2652
 Number of obs: 10000, groups: person, 1000; item, 10

> itemest_ = coef(raschRnd)$item[, 1]                      # extracting item estimates
> itemRnd=matrix(itemest_)
```

(continued)

TABLE 13.17. (continued)

```

> itemRnd
      [,1]
[1,]  2.1298815
[2,]  1.5045406
[3,]  1.1104772
[4,]  0.4902315
[5,]  0.1528053
[6,] -0.2073199
[7,] -0.3188731
[8,] -0.7510732
[9,] -1.3770191
[10,] -1.8761206

> raschRnd_thetahat = coef(raschRnd)$person[, 1] # extracting people estimates
> peopleRnd=matrix(raschRnd_thetahat)
> head( peopleRnd,5)
      [,1]
[1,] -0.30306342
[2,] -0.30306342
[3,] -0.30306342
[4,] -0.30306342
[5,] -0.03873654
> tail(peopleRnd,n=2)
      [,1]
[999,]  0.2255766
[1000,] -0.5695346

> # LLTM+e (item random effects & fixed components)
> LLTMe=glmer(x~0+(1|item) + q1 + q2 + (1|person), family=binomial("logit"),
  data=stacked, control=optmzrinfo)
> summary(LLTMe)
Generalized linear mixed model fit by maximum likelihood
  (Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + (1 | item) + q1 + q2 + (1 | item) + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC      logLik deviance df.resid
 11225.9  11254.8   -5609.0   11217.9     9996

Scaled residuals:
    Min      1Q  Median      3Q      Max
-4.2527 -0.6769  0.2693  0.6692  3.8848

Random effects:
 Groups Name        Variance Std.Dev.
 person (Intercept) 0.5295   0.7277
 item   (Intercept) 0.4810   0.6935
 Number of obs: 10000, groups: person, 1000; item, 10

Fixed effects:
 Estimate Std. Error z value Pr(>|z|)
 Q1    1.5906     0.3494   4.552 5.32e-06 ***
 Q2   -1.1620     0.3027  -3.838 0.000124 ***
 ---

```

(continued)

TABLE 13.17. (continued)

```

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
  Q1
Q2 -0.575

> anova(raschRnd, raschLLTMe)
Data: stacked
Models:
raschRnd: x ~ 0 + (1 | item) + (1 | person)
raschLLTMe: x ~ 0 + (1 | item) + Q1 + Q2 + +(1 | person)
  Df AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
raschRnd    2 11234 11248 -5614.9     11230
raschLLTMe  4 11226 11255 -5609.0     11218 11.951      2  0.002541 **
---
Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '.' 0.1 ' ' 1

> itemest_ = coef(LLTMe)$item[, 1]                                # extracting item estimates
> itemLLTMe=matrix(itemest_)
> itemLLTMe
 [,1]
[1,]  0.54103764
[2,] -0.07900892
[3,]  0.67764644
[4,]  0.06207209
[5,] -0.27332588
[6,]  0.94532111
[7,]  0.83434899
[8,] -1.17069371
[9,] -0.21797040
[10,] -0.71195580

> raschLLTMe_thetahat = coef(LLTMe)$person[, 1]                  # extracting person estimates
> peopleLLTMe=matrix(raschLLTMe_thetahat)
> head(peopleLLTMe,5)
 [,1]
[1,] -0.3027268
[2,] -0.3027268
[3,] -0.3027268
[4,] -0.3027268
[5,] -0.0383219

> tail(peopleLLTMe,2)
 [,1]
[999,]  0.2260454
[1000,] -0.5692985

```

13.11 except for the addition of the g subscript to reflect the nesting variable (e.g., ζ_{00g} , θ_{0ig}^r , ζ_{q0g}). Accordingly, we have

$$\begin{aligned}\beta_{1ig} &= \zeta_{10g} \\ &\vdots\end{aligned}$$

$$\beta_{(L-1)ig} = \varsigma_{(L-1)0g}$$

$$\beta_{0ig} = \varsigma_{00g} + \theta_{0ig}^r. \quad (13.34)$$

In Equation 13.34, θ_{0ig} reflects the deviation of person i from mean of the persons in group g with within-group variance σ_ζ^2 . θ_{0ig} is assumed to be $N(0, \sigma_\zeta^2)$, with σ_ζ^2 assumed to be constant across the nesting variable.

At level-3 (groups) we predict our level-2 ς_{q0g} s from our level-3 information

$$\begin{aligned}\varsigma_{00g} &= \pi_{000} + \varepsilon_{00g}^r \\ \varsigma_{10g} &= \pi_{100} \\ \varsigma_{20g} &= \pi_{200} \\ &\vdots \\ \varsigma_{(L-1)0g} &= \pi_{(L-1)00},\end{aligned}\quad (13.35)$$

where ε_{00g}^r is a random effect of ς_{00g} assumed to be normally distributed with variance σ_π^2 and mean 0; the π s are fixed effects. Analogous to the two-level model, the item location for the L th item is given by the fixed effects parameter (i.e., π_{000}) associated with the constant term, ς_{00g} , with the remaining ($L - 1$) item locations given by the deviations from π_{000} . That is, $\delta_1 = (-\pi_{100} - \pi_{000})$, $\delta_2 = (-\pi_{200} - \pi_{000})$, and so on.

By substituting Equations 13.34 and 13.35 into Equation 13.33, a single model comprising all three levels is obtained:

$$\begin{aligned}\vartheta_{ijg} &= \beta_{0ig} + \beta_{1ig} X_{1ig} + \beta_{2ig} X_{2ig} + \dots + \beta_{(L-1)ig} X_{(L-1)ig}, \\ &= (\varsigma_{00g} + \varepsilon_{00g}^r) + \varsigma_{10g} + \dots + \varsigma_{(L-1)0g} \\ &= (\pi_{000} + \varepsilon_{00g}^r) + \theta_{0ig}^r + \pi_{100} + \dots + \pi_{(L-1)00}.\end{aligned}\quad (13.36)$$

Therefore, by substitution, for item j we have

$$P_{jis} = \frac{\exp[\vartheta_{ijg}]}{1 + \exp[\vartheta_{ijg}]} = \frac{\exp[\theta_{0ig}^r - \delta_j]}{1 + \exp[\theta_{0ig}^r - \delta_j]}, \quad (13.37)$$

where $\delta_j = (-\pi_{j00} - \pi_{000})$ and person i 's ability in group g , is $\theta_{0ig} = \varepsilon_{00g}^r + \theta_{0ig}^r$. Analogous to above, we have a single model that includes information about item variability, person variability, and group variability. In short, our model accounts for the dependencies that exist due to the nesting of individuals within groups. These basic ideas also allow for developing models for longitudinal assessment (e.g., Bianconcini & Casgno, 2010), studying conditional item dependence (Jiao, Wang, & Kamata, 2005), as well as

incorporating latent classes (e.g., Cho & Cohen, 2010; Vermunt, 2003, 2007), to name just a few uses.

Example: Three-Level Model Analysis— Nutrition Literacy, proc glimmix

For our example, we revisit our Person-Level Predictors for Respondents example in which we introduced an assessment scale designed to measure nutrition literacy. Recall that this scale is used to assess the effectiveness of a nutrition education program. This program's goal is to improve nutrition literacy and thereby reduce health disparities, as well as improve preventive care and health outcomes. The program is offered in a hybrid face-to-face/online format through the county-based extension offices of 20 Land Grant universities, with the office's nutrition extension educator as the instructor of record. These county-based extension educators working out of county-based extension offices are our level-3. Program attendees were individuals primarily responsible for purchasing a household's groceries and cooking meals. In total, there were 1000 respondents evenly distributed across the 20 extension educators. Thus, level-1 is our attendees' responses, level-2 is our attendees, and level-3 is our extension educators.

Table 13.18 contains our `glimmix` program; the data steps from Table 13.1 (not shown) are used to create the stacked _ data set from the file NutrLit.dat. As can be seen, we treat our items as fixed and our respondents as well as educators as random. Consequently, we have a `random` statement for our respondents and one for our educators. It is in the `random` statement that we indicate that our level-2 respondents (i.e., person) are nested within our level-3 educators (i.e., `person(educator)`). Additionally, we now use `covtest` to obtain Wald z-tests for our covariance parameter estimates.

Table 13.19 contains our output. Comparing our three-level Fit Statistics with those in which we ignore level 3 (i.e., a two-level analysis; Table 13.19 bottom panel) shows that each of our information criteria is slightly smaller with our three-level analysis than with the two-level model. Therefore, including a third level produces a better model-data fit than ignoring the nesting due to educator. Our covariance tests for respondents nested within educator and educator are significant at the 5% significance

TABLE 13.18. proc glimmix Nutrition Literacy—3-Level

```
proc glimmix data=stacked_data noclprint method=laplace;
  title "fixed items & random P & random educator; 3 level; method=laplace";
  class item person educator;
  model x(event='1')= item / dist=binary link=logit solution;
  random intercept / subject=educator solution;
  random intercept / subject=person(educator) solution;
  covtest / wald;
  lsmeans item / ilink;
run;
```

TABLE 13.19. proc glimmix Nutrition Literacy—3- and 2-Level Output

Fit Statistics					
-2 Log Likelihood					10962.46
AIC (smaller is better)					10986.46
AICC (smaller is better)					10986.49
BIC (smaller is better)					10998.41
CAIC (smaller is better)					11010.41
HQIC (smaller is better)					10988.79
Fit Statistics for Conditional Distribution					
-2 log L(x r. effects)					10117.44
Pearson Chi-Square					8994.51
Pearson Chi-Square / DF					0.90
Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	educator	0.2493	0.08481	2.94	0.0016
Intercept	person(educator)	0.3035	0.04328	7.01	<.0001
Solutions for Fixed Effects					
Effect	item	Estimate	Standard Error	DF	t Value
Intercept		-1.9884	0.1478	19	-13.45 <.0001
item	1	4.2387	0.1428	8991	29.68 <.0001
item	2	3.5854	0.1292	8991	27.75 <.0001
item	3	3.1713	0.1237	8991	25.64 <.0001
item	4	2.5146	0.1184	8991	21.24 <.0001
item	5	2.1553	0.1170	8991	18.42 <.0001
item	6	1.7711	0.1166	8991	15.19 <.0001
item	7	1.6521	0.1167	8991	14.16 <.0001
item	8	1.1912	0.1182	8991	10.08 <.0001
item	9	0.5269	0.1240	8991	4.25 <.0001
item	10	0	.	.	.
Solution for Random Effects					
Effect	Subject	Estimate	Std Err Pred	DF	t Value
Intercept	educator 1	-0.2519	0.1663	8991	-1.51 0.1299
Intercept	person(educator) 1 1	-0.1342	0.4444	8991	-0.30 0.7627
Intercept	person(educator) 2 1	-0.1342	0.4444	8991	-0.30 0.7627
Intercept	person(educator) 3 1	-0.1342	0.4444	8991	-0.30 0.7627
Intercept	person(educator) 4 1	-0.1342	0.4444	8991	-0.30 0.7627
Intercept	person(educator) 5 1	0.06066	0.4434	8991	0.14 0.8912
:					
Intercept	person(educator) 50 1	-0.1342	0.4444	8991	-0.30 0.7627
Intercept	educator 2	0.1160	0.1663	8991	0.70 0.4855
Intercept	person(educator) 51 2	0.1230	0.4440	8991	0.28 0.7818
Intercept	person(educator) 52 2	-0.2660	0.4440	8991	-0.60 0.5491
Intercept	person(educator) 53 2	0.1230	0.4440	8991	0.28 0.7818
Intercept	person(educator) 54 2	-0.4607	0.4460	8991	-1.03 0.3017
Intercept	person(educator) 55 2	-0.07164	0.4433	8991	-0.16 0.8716
Intercept	person(educator) 56 2	0.1230	0.4440	8991	0.28 0.7818
:					
Intercept	person(educator) 997 20	0.2663	0.4439	8991	0.60 0.5485
Intercept	person(educator) 998 20	-0.3190	0.4468	8991	-0.71 0.4752
Intercept	person(educator) 999 20	0.2663	0.4439	8991	0.60 0.5485
Intercept	person(educator) 1000 20	-0.3190	0.4468	8991	-0.71 0.4752

(continued)

TABLE 13.19. (continued)

item Least Squares Means						
item	Estimate	Standard Error	DF	t Value	Pr > t	Standard Error Mean
1	2.2503	0.1533	8991	14.68	<.0001	0.9047
2	1.5970	0.1411	8991	11.32	<.0001	0.8316
3	1.1830	0.1364	8991	8.67	<.0001	0.7655
4	0.5262	0.1324	8991	3.98	<.0001	0.6286
5	0.1669	0.1315	8991	1.27	0.2045	0.5416
6	-0.2172	0.1316	8991	-1.65	0.0989	0.4459
7	-0.3363	0.1319	8991	-2.55	0.0108	0.4167
8	-0.7971	0.1337	8991	-5.96	<.0001	0.3106
9	-1.4615	0.1396	8991	-10.47	<.0001	0.1882
10	-1.9884	0.1478	8991	-13.45	<.0001	0.1204
						0.01566

Two-level analysis						
•						
Fit Statistics						
-2 Log Likelihood						
11158.59						
AIC (smaller is better)						
11180.59						
AICC (smaller is better)						
11180.62						
BIC (smaller is better)						
11234.58						
CAIC (smaller is better)						
11245.58						
HQIC (smaller is better)						
11201.11						
•						
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error			
Intercept	person	0.5463	0.05628			
•						
Solution for Random Effects						
Effect	Subject	Estimate	Std Err Pred	DF	t Value	Pr > t
Intercept	person 1	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 2	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 3	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 4	-0.3128	0.5229	8991	-0.60	0.5497
Intercept	person 5	-0.04072	0.5215	8991	-0.08	0.9378
•						
Intercept	person 997	0.2313	0.5226	8991	0.44	0.6581
Intercept	person 998	-0.5871	0.5267	8991	-1.11	0.2650
Intercept	person 999	0.2313	0.5226	8991	0.44	0.6581
Intercept	person 1000	-0.5871	0.5267	8991	-1.11	0.2650
•						
item Least Squares Means						
item	Estimate	Standard Error	DF	t Value	Pr > t	Standard Error Mean
1	2.2517	0.1062	8991	21.19	<.0001	0.9048
2	1.5975	0.08780	8991	18.19	<.0001	0.8317
3	1.1825	0.08005	8991	14.77	<.0001	0.7654
4	0.5239	0.07288	8991	7.19	<.0001	0.6281
5	0.1636	0.07134	8991	2.29	0.0218	0.5408
6	-0.2213	0.07149	8991	-3.10	0.0020	0.4449
7	-0.3405	0.07190	8991	-4.74	<.0001	0.4157
8	-0.8009	0.07521	8991	-10.65	<.0001	0.3098
9	-1.4627	0.08504	8991	-17.20	<.0001	0.1880
10	-1.9860	0.09780	8991	-20.31	<.0001	0.1207
						0.01038

level; the respondent variability of 0.5463 with the 2-level analysis is reduced to 0.3035 by taking into consideration educators. In short, differences in educator instructional quality significantly account for respondent differences in nutritional literacy.

As seen from the Item Least Squares Means table, our 3- and 2-level analyses produce, as expected, essentially the same estimated item locations. The correlation between the two sets of estimates is 0.999998 with Ms for the 2- and 3-level of 0.0908 ($SD = 1.2700$) and .0923 ($SD = 1.2696$), respectively.

The Solution for Random Effects table contains the respondent and educator estimated locations intermixed. As expected given our use of the Rasch model, our first four respondents each with $X = 4$ are estimated to have the same location on the nutritional literacy continuum ($\hat{\theta}_{0,1,1}^r = -0.1342$) with an estimated educator effect of -0.2519; the first 50 respondents were instructed by the same educator, and thus all have the same educator effect. Further down through the table, we find our second educator and their corresponding respondents (i.e., person(educator) 51 2, person(educator) 52 2; etc.). Our second educator's estimated effect is larger than that of the first one ($\hat{\varepsilon}_{002}^r = 0.1160$), and the 52nd respondent ($\hat{\theta}_{0,52,2}^r = 0.1230$, $X = 4$) is estimated to have more nutritional literacy than the first four respondents with the same X. Comparing our 3-level $\hat{\theta}_{0ig}^r$ s and 2-level $\hat{\theta}_{ig}^r$ s shows that they are different. However, the two are correlated 0.89. When we take into account the educator contribution ($\hat{\varepsilon}_{00g}^r$) to $\hat{\theta}_{ig}^r$ (i.e., $\hat{\theta}_{ig}^r = \hat{\varepsilon}_{00g}^r + \hat{\theta}_{0ig}^r$), the two sets show an average difference of -0.0005.

Example: Three-Level Analysis of Nutrition Literacy Data, lme4

We now reanalyze our data using `lme4`. Table 13.20 contains our R session. In lieu of `tidyR`, we now use `reshape2` (Wickham, 2007) for transforming our unstacked data to stacked data. As is the case with `tidyR`, this transformation is a two-step process involving `reshape2`'s `melt` function and R's `order` function (these functions are analogous to `gather` and `arrange`, respectively). We first use the `melt` function to transform our unstacked data into a stacked format and store the result in the `tmpstacked` data frame. The first argument to `melt` is the unstacked data frame, followed by a list of variables that are *not* to be nested (`id.vars = c("person", ...)`), the variables to be considered measured (i.e., to be nested), and the variable names to be given to the nested variables (`variable.name = "item," value.name = "x"`). We then use the `order` function to reorder the rows so that all the item responses for a respondent are contiguous and conform to Figure 13.2. (For pedagogical reasons, we use `tmpstacked`. However, both the `melt` and `order` functions could be called using `stacked` in lieu of `tmpstacked` (i.e., `stacked = melt(unstacked, id.vars = ... "x")` and `stacked = with(stacked, stacked[order(educator, ... ,)])`)).

Our first analysis fits the Rasch model (`rasch = glmer(x~0 + item + (1|person) ...)`) to our three-level data. As can be seen, our person variability, 0.5457, is slightly less than the `glimmix` estimate. Our `glmer` and `glimmix` item location and person location estimates also show almost perfect agreement.

To perform the three-level analysis, we introduce the educator random effect into

TABLE 13.20. lme4 Session for Nutrition Literacy 3-Level Example

```

> library(reshape2); library(lme4); library(optimx)

> unstacked = read.table(file.choose(), header=F) # Read data file NutrLit.dat

> # create meaningful variable names
> names(unstacked)=c('person','PwiEdctr','educator','edlevel','i1','i2','i3', 'i4',
+ 'i5', 'i6','i7','i8','i9','i10')

> # reformat unstacked data into stacked format using reshape2
> tmpstacked=melt(unstacked, id.vars =c("person","edlevel","educator","PwiEdctr"),
+ measure.vars=c('i1','i2','i3','i4','i5','i6','i7','i8','i9','i10'),
+ variable.name="item",value.name="x")
# reorder data to be items w/i PwiEdctr w/i educator
> stacked=with(tmpstacked,tmpstacked[order(educator,PwiEdctr,item),])

> head(stacked,12)
   person edlevel educator PwiEdctr item   x
1       1     18        1       1   i1   1
1001    1     18        1       1   i2   1
2001    1     18        1       1   i3   0
3001    1     18        1       1   i4   1
4001    1     18        1       1   i5   0
5001    1     18        1       1   i6   0
6001    1     18        1       1   i7   1
7001    1     18        1       1   i8   0
8001    1     18        1       1   i9   0
9001    1     18        1       1  i10  0
2       2     18        1       2   i1   1
1002    2     18        1       2   i2   1

># to speed up execution optmzrinfo - see Table 13.5

> # doing the rasch model w/o intercept (2-level)
> rasch=glmer(x~0+item+(1|person),family=binomial("logit"),data=stacked,
+ control=optmzrinfo)

> summary(rasch)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + item + (1 | person)
Data: stacked
Control: optmzrinfo

      AIC      BIC  logLik deviance df.resid
11180.6  11259.9 -5579.3  11158.6      9989

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.5615 -0.6676  0.2520  0.6544  4.1605

Random effects:
Groups Name        Variance Std.Dev.
person (Intercept) 0.5457   0.7387
Number of obs: 10000, groups: person, 1000

```

(continued)

TABLE 13.20. (*continued*)

```

Fixed effects:
    Estimate Std. Error z value Pr(>|z|)
itemi1  2.25156   0.10749  20.948 < 2e-16 ***
itemi2  1.59728   0.08718  18.322 < 2e-16 ***
itemi3  1.18227   0.07874  15.015 < 2e-16 ***
itemi4  0.52378   0.07105  7.372  1.68e-13 ***
itemi5  0.16357   0.06943  2.356  0.01848 *
itemi6 -0.22127   0.06960 -3.179  0.00148 **
itemi7 -0.34040   0.07004 -4.860  1.17e-06 ***
itemi8 -0.80080   0.07357 -10.885 < 2e-16 ***
itemi9 -1.46260   0.08419 -17.372 < 2e-16 ***
itemi10 -1.98590  0.09818 -20.227 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      itemi1 itemi2 itemi3 itemi4 itemi5 itemi6 itemi7 itemi8 itemi9
itemi2  0.062
itemi3  0.068  0.083
itemi4  0.073  0.090  0.099
itemi5  0.074  0.091  0.100  0.111
itemi6  0.072  0.089  0.099  0.110  0.113
itemi7  0.071  0.088  0.098  0.109  0.112  0.112
itemi8  0.067  0.083  0.092  0.103  0.107  0.107  0.107
itemi9  0.057  0.071  0.079  0.090  0.093  0.094  0.094  0.090
itemi10 0.048  0.060  0.067  0.076  0.079  0.080  0.080  0.078  0.070

> thetaestRasch=coef(rasch)$person[,1]                                     # extracting respondent est
> peopleRasch=matrix(thetaestRasch)
[1]
[1,] -0.31261922
[2,] -0.31261922
[3,] -0.31261922
[4,] -0.31261922
[5,] -0.04068828
:
[997,]  0.23117456
[998,] -0.58678953
[999,]  0.23117456
[1000,] -0.58678953

> # 3-level analysis with respondent nested within educator
> rasch3L=glmer(x~0+item+(1|person)+(1|educator),family=binomial("logit"),
  data=stacked,control=optmzrinfo)

> anova(rasch,rasch3L)
Data: stacked
Models:
rasch: x ~ 0 + item + (1 | person)
rasch3L: x ~ 0 + item + (1 | person) + (1 | educator)
          Df AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
rasch   11 11181 11260 -5579.3     11159
rasch3L 12 10986 11073 -5481.2     10962 196.14      1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(continued)

TABLE 13.20. (continued)

```

> summary(rasch3L)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + item + (1 | person) + (1 | educator)
Data: stacked
Control: ctrlparm

      AIC      BIC   logLik deviance df.resid
10986.5 11073.0 -5481.2 10962.5     9988

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.5806 -0.6706  0.2301  0.6554  4.6771

Random effects:
Groups   Name        Variance Std.Dev.
person   (Intercept) 0.3034   0.5508
educator (Intercept) 0.2491   0.4991
Number of obs: 10000, groups: person, 1000; educator, 20

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
itemi1     2.2502    0.1537 14.637 < 2e-16 ***
itemi2     1.5968    0.1407 11.349 < 2e-16 ***
itemi3     1.1827    0.1358  8.711 < 2e-16 ***
itemi4     0.5261    0.1316  3.998 6.39e-05 ***
itemi5     0.1669    0.1308  1.276  0.2019
itemi6    -0.2173    0.1309 -1.660  0.0969 .
itemi7    -0.3363    0.1311 -2.565  0.0103 *
itemi8    -0.7970    0.1330 -5.992 2.07e-09 ***
itemi9    -1.4615    0.1391 -10.506 < 2e-16 ***
itemi10   -1.9883    0.1479 -13.448 < 2e-16 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Correlation of Fixed Effects:
          itemi1 itemi2 itemi3 itemi4 itemi5 itemi6 itemi7 itemi8 itemi9
itemi2  0.593
itemi3  0.614  0.670
itemi4  0.632  0.690  0.715
itemi5  0.635  0.694  0.719  0.742
itemi6  0.634  0.693  0.718  0.741  0.746
itemi7  0.632  0.691  0.716  0.740  0.744  0.744
itemi8  0.622  0.680  0.705  0.728  0.733  0.733  0.732
itemi9  0.594  0.649  0.673  0.696  0.701  0.701  0.700  0.691
itemi10 0.558  0.610  0.633  0.654  0.659  0.660  0.659  0.650  0.623

> rasch3L_educatoreff = coef(rasch3L)$educator[, 1]           # extracting educator est
> EducatorEst=matrix(data=rasch3L_educatoreff)                 # reformatting educator estimates
> EducatorEst
      [,1]
[1,] -0.251848446
[2,]  0.116078051
[3,]  0.250588912

```

(continued)

TABLE 13.20. (continued)

```
[4,] -0.563922518
[5,] -0.733398056
:
[17,]  0.501715561
[18,] -0.755305214
[19,]  0.363814100
[20,] -0.283163207

> rasch3L_thetahat = coef(rasch3L)$person[, 1] # extracting respondent est
> people=matrix(rasch3L_thetahat)

> head(people,5)
      [,1]
[1,] -0.13415557
[2,] -0.13415557
[3,] -0.13415557
[4,] -0.13415557
[5,]  0.06064093

> people[50:56,]
 [1] -0.1341556  0.1229463 -0.2659337  0.1229463 -0.4605863 -0.0716143  0.1229463

> tail(people,4)
      [,1]
[997,]  0.2662439
[998,] -0.3189419
[999,]  0.2662439
[1000,] -0.3189419
```

our model ($\text{rasch3L} = \text{glmer}(x \sim 0 + \text{item} + (1|\text{person}) + (1|\text{educator}), \dots)$). Comparing our information criteria, we see that both AIC and BIC indicate better model-data fit with the three-level model than with the two-level model. Furthermore, ΔG^2 also indicates a significant improvement in model–data fit by the three-level model relative to the Rasch model ($\text{anova}(\text{rasch}, \text{rasch3L})$). As is the case above, our person variability estimate is reduced to 0.3034 by introducing our educator effect into the model. Our education effects (EducatorEst) agree with those of glimmix. The first educator's estimated effect ($\hat{\varepsilon}_{001}^r$) for the first 50 respondents is -0.251848446 , the second educator's estimated effect ($\hat{\varepsilon}_{002}^r$) for the second 50 respondents is 0.116078051 , and so forth. Our first respondent's nutrition literacy estimated location ($\hat{\theta}_{0,1,1}^r$) given the first educator's effect is -0.13415557 , whereas the fifty-first respondent's nutrition literacy estimated location ($\hat{\theta}_{0,51,2}^r$), given the second educator's effect, is 0.1229463 , etc. As above, our 3-level $\hat{\theta}_{0ig}^r$ s and 2-level $\hat{\theta}_{ig}$ s are different. However, the two show a strong linear relationship of 0.89. When we take into account the educators' contribution ($\hat{\varepsilon}_{00g}^r$) to $\hat{\theta}_{ig}$ (i.e., $\hat{\theta}_{ig} = \hat{\varepsilon}_{00g}^r + \hat{\theta}_{0ig}^r$), the two sets show an average difference of 0.0000. Our item location estimates (e.g., $\hat{\delta}_1^E = 2.2502$, $\hat{\delta}_2^E = 1.5968$, $\hat{\delta}_3^E = 1.1827$) are perfectly correlated with those of glimmix.

A slightly more complicated example would involve adding level-3 predictors of the

differences in our educators. For instance, assuming educators had a significant impact, we might want to account for the variability in our educators. To do this, we could add one or more fixed effects predictors such as educator experience and/or an instructor quality measure to account for differences in our educators' instructional ability.

Summary

The models presented in Chapters 1–10 can be conceptualized from a multilevel perspective. The advantages of this perspective are an increased precision with which factors' effects are estimated, the simultaneous estimation of item/person/predictor effects at multiple levels, and the ability to correctly account for the hierarchical nature of some data sets (see Pastor, 2003; Sulis & Toland, 2017).

The reconceptualization of an IRT model such as the Rasch model as a multilevel model requires that we have two levels—an item level and a person level. At level-1 are the item responses, whereas at level-2 we have persons. Hierarchically, item responses are nested within persons. With multilevel modeling, it is necessary to distinguish between fixed effects and random effects predictors. Random effects predictors have distributional assumptions (e.g., $\theta_{01}^r \sim N(0, \sigma_\theta^2)$), whereas fixed effects predictors do not. In short, the fundamental distinction is whether a predictor's values are held constant (fixed) or whether they are allowed to vary (random). Because in the standard Rasch model items are considered to have the same effect for each person, item effects are modeled as fixed effects. In contrast, because person locations are assumed to vary across persons, they are treated as random effects. Therefore, the modeling of persons consists of a constant effect (the intercept) plus a random residual effect. The residual after fitting the constant at level-2 is the person's location estimate.

In addition to persons, level-2 can also include item- and/or person-oriented predictors that are believed to affect the observed item responses. The use of these predictors in our models is the primary benefit of adopting a multilevel perspective.

Some data include a third nesting level such as country (i.e., level-1: responses, level-2: persons, level-3: country). These data are fitted using a 3-level model. As is the case with 2-level models, one needs to decide on whether an effect should be treated as fixed or random. With these data the analysis would first determine the necessity of accounting for the third level. Assuming that it is necessary to account for the third level, our 3-level models will correctly attribute the relative importance of all influential sources on the item responses. Failure to account for the third level will most likely lead to inaccurate interpretations. A three-level model may or may not contain item- and/or person-oriented predictors at level 2.

Notes

1. Although our treatment has items nested within people, Adams and Wilson (1996) used the terms *items* and *cases* to indicate that a case "is any object to be measured"

(p. 145) and not necessarily a person. As an example, Mislevy (1983) defines an IRT model at the group or subpopulation level rather than at the individual level.

2. Multilevel models are also called hierarchical linear models (HLMs; Raudenbush & Bryk, 2002); general linear mixed models (Rabe-Hesketh & Skrondal, 2002); hierarchical generalized linear mixed models; multilevel generalized linear mixed models; generalized linear mixed (effects) models (Kamata, 1998; Skrondal & Rabe-Hesketh, 2009); generalized linear mixed models (McCullagh & Searle, 2001); random coefficients multinomial logit models (Adams & Wilson, 1996; Adams, Wilson, & Wu, 1997); and sometimes random coefficient models (de Leeuw & Kreft, 1986). Generalized linear mixed models subsume IRT, latent class, factor analytic, and structural equation models.
3. Multilevel IRT is not limited to the Rasch model (see Van den Noortgate, De Boeck, & Meulders, 2003). For example, Fox (2004) uses multilevel IRT modeling with the 2P normal ogive model; Bacci and Caviezel (2011) use multilevel modeling with the generalized partial credit model; van Nispen, Knol, Neve, and van Rens (2010) use a multilevel approach with the GR model. and Huang (2015) used a three-parameter model.
4. Both generalized linear models and generalized linear mixed models (GLMM) are used with a non-normally distributed response (outcome) variable. However, the former is restricted to when one has fixed effects predictors, whereas GLMM can be used when one has both random and fixed effects predictors of this response variable. To use the GLMM framework, it is necessary to have a probability distribution for the response variable given one's expectation and a link function that relates the linear predictor model (of random and/or fixed effects predictors) to the expectation of the response variable. The link function "linearizes" the relationship between the response variable and the predictors. The probability distribution's mean and variance are also required for the implementation. Various probability distributions for the response variable are the Bernoulli (or, more generally, the binomial), Poisson, normal, multinomial, and so on. Possible link functions include the probit, log, log-log, identity, and logit link functions, to name a few. In the case of the Rasch model, the link function is the logit or logistic link. In the case of the Bernoulli distribution, the mean of our item response conditional on the probability of a response of 1 is p_{ij} with variance $p_{ij}(1 - p_{ij})$; that is, $\epsilon(x_{ij}|p_{ij}) = p_{ij}$ and $Var(x_{ij}|p_{ij}) = p_{ij}(1 - p_{ij})$. With our GLMM we are, in effect, trying to predict the mean of the response variable as a function of a set of predictors.
5. The predictors (X s) can be collected into a *design matrix*, \underline{X} . This is an item by predictor matrix in which columns 2 through L are indicator variables (i.e., predictors) and column 1 reflects the constant, β_{0j} ; some prefer the term *dummy variables* to *indicator variables*. The entries in the matrix reflect the presence (cell contains a 1) or absence (cell contains a 0) of an item. Equation 13.3 is for person i . Thus, we have L equations for person i , and in matrix form we have

$${}_L \underline{\boldsymbol{\theta}}_1 = {}_L \underline{\mathbf{X}} {}_L \underline{\mathbf{b}}_1 \quad (13.38)$$

$$\begin{bmatrix} \vartheta_{i1} \\ \vartheta_{i2} \\ \vdots \\ \vartheta_{i(L-1)} \\ \vartheta_{iL} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \ddots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \\ \beta_{2i} \\ \vdots \\ \beta_{Li} \end{bmatrix} \quad (13.39)$$

For $\underline{\mathbf{X}}$ to be of full rank, we have $L - 1$ indicator variables (i.e., $\beta_{Li} = 0$). As such, the entries in the last row of $\underline{\mathbf{X}}$ are 1 in column 1 (the constant) and 0 in columns 2 through L.

6. $\sum_{q=1}^{L-1} \beta_{qi} X_{qij}$ simplifies to β_{qi} because

$$\beta_{qi} X_{qij} = \beta_{qi}(1)$$

when $q = j$, and when $q \neq j$

$$\beta_{qi} X_{qij} = \beta_{qi}(0) = 0.$$

7. The error associated with Equation 13.7 is absorbed by the link function (O'Connell, Goldstein, Rogers, & Peng, 2008).
8. If the item is located at 1.5 for some respondents (e.g., males), but at -1.0 for other respondents (e.g., females), then we have evidence of differential item function.
9. The Rasch model treats items and persons as fixed effects.
10. SAS proc nlmixed could also be used. nlmixed uses numerical integration for estimation, whereas glimmix uses a Taylor series expansion. The former is more computationally and time intensive than the latter, but is potentially more accurate than glimmix. Additionally, glimmix can handle more random effects than nlmixed. However, the criterion for glimmix must be from an exponential distribution, whereas nlmixed is more flexible in this regard (Flom, McMahon, & Pouget, 2007).
11. To convert an unstacked data set to a stacked format in SPSS, one would use the Data menu's Restructure . . . item.
12. If we wanted the Fixed Effects table, we would add solution to the model statement: `model x(event = '1') = item / dist = binary link = logit solution;`
13. For comparison purposes, a base model can be estimated in which we assume that all items have the same location and responses are not nested within persons (i.e., there are no predictors of item variability except the mean). The glimmix program for estimating this model is

```
proc glimmix data = d1 method = quadrature;
  title "All items have same location";
  model x(event = '1') = / dist = binary link = logit solution;
```

The corresponding output is

Fit Statistics	
-2 Log Likelihood	6930.89
AIC (smaller is better)	6932.89
AICC (smaller is better)	6932.89
BIC (smaller is better)	6939.41
CAIC (smaller is better)	6940.41
HQIC (smaller is better)	6935.17
Pearson Chi-Square	5000.00
Pearson Chi-Square / DF	1.00

Parameter Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	0.02160	0.02829	4999	0.76	0.4451

The model-data fit is worse than when we include items as fixed effects and persons as random effects in the model (Table 13.2). On average, the probability of a response of 1 on any item is

$$P_{ij} = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{0.0216}}{1+e^{0.0216}} = 0.5054$$

14. One can avoid performing calculations to obtain $\hat{\delta}_j^E$ and its standard error by running a no intercept model and including *all* the indicators. For example,

```
proc glimmix data = d1 method = quadrature;
title "MM-Rasch formulation demo, design matrix, noncalc ex";
model x(event = 'x1') = x1 x2 x3 x4 x5/noint solution dist = binary
link = logit;
random intercept/subject=person G solution;
output out=fit _statistics resid(ilink)=residual variance(ilink)=
variance;
```

As can be seen, the fixed effects coefficients and their corresponding standard errors correspond to those found in the Item Least Squares Means table.

15. Pearson's chi-square is Yen (1981) Q_1 statistic and the Likelihood Ratio statistic, G^2 , is the one proposed by McKinley and Mills (1985). To calculate each statistic,

Solutions for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
x1	2.1303	0.1044	3996	20.40	<.0001
x2	0.2745	0.06705	3996	4.09	<.0001
x3	-0.04207	0.06644	3996	-0.63	0.5267
x4	-0.8095	0.07185	3996	-11.27	<.0001
x5	-1.0413	0.07537	3996	-13.82	<.0001

our respondents are allocated to 10 fractiles of approximately equal size based on their $\hat{\theta}$ s and the observed and expected proportions for each calculated. These statistical tests are only available for dichotomous data. Retrieved from https://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htm#statug_irt_details13.htm September 17, 2018.

16. With a small number of items, the creation of 10 fractiles may be problematic. For example, with five items there are only six Xs and corresponding $\hat{\theta}$ s. Categorizing the continuum into more than six fractiles of approximately equal size can be problematic given the θ distribution's shape.
17. This analysis was run on a 64-bit system (SAS 9.4 X64_10Pro). On a 32-bit system (SAS 9.4 W32_7Pro), the various estimates, item fit statistics, G1–G4 means had different values than presented here. In some cases, the discrepancies were in the fifth or sixth decimal value, whereas in other cases (e.g., item fit statistics) the discrepancies were substantial. Moreover, on the 32-bit system it was observed that not all examinees with the same X received the same $\hat{\theta}$. For example, not all individuals with a X = 3 receive the same $\hat{\theta}$. Specifically, only response patterns of 1110, 10110, 01110, 01011 received a $\hat{\theta} = 0.11195$. However, response patterns of 11001, 10110, 10101, 11010, and 01101 received a $\hat{\theta} = 0.11196$, and the response pattern of 00111 received a $\hat{\theta} = 0.11191$. Nevertheless, these discrepancies in the fifth decimal place are not practically significant.
18. If the data contain zero and perfect scores, then the estimates differ ever so slightly on a 64-bit machine. For example, `proc glimmix` and `proc irt estimate`, respectively, item 1 to be at 2.4899 and -2.49096, item 2 to be at 0.7342 and -0.73449, and so on; the `glimmix` average $\hat{\delta} = 0.50034$ ($SD = 1.11961$) and the `proc irt` average $\hat{\delta} = -0.50069$ ($SD = 1.11966$). However, the two sets of item location estimates $r = -1.0$.
19. One can make the transformation from unstacked to stacked format by combining the two steps and using the redirection operator (i.e., piping) “%>%”:

```
stacked = unstacked %>% gather(key = item,value =
  x,i1:i5) %>% arrange(person,item)
```

That is, `unstacked %>% gather(key = ...)` indicates that the unstacked data frame is to be sent (i.e., piped) to the `gather` function and its results are to be sent to the `arrange` function (i.e., `%>% arrange(person, ...)`).

20. We are following the anonymously authored performance tips provided at <https://cran.r-project.org/web/packages/lme4/vignettes/lmerperf.html>.
21. To convert the Fixed Effects to those in the Item*race Least Squares Means table, we need to take into account the intercept as well as the race, item, and item-race interaction effects (as appropriate). For the Focal ($\hat{\delta}_{j,F}^E$) and Reference ($\hat{\delta}_{j,R}^E$) groups we have:

$$\hat{\delta}_{j,F}^E = \text{Intercept} + (\text{item}_j \text{ effect}) + (\text{race_F effect})$$

$$\hat{\delta}_{j,R}^E = \text{Intercept} + (\text{item}_j \text{ effect}) + (\text{race_R effect}) + (\text{item}_j * \text{race_R effect})$$

where `race_F` and `race_R` effects reflect the Focal and Reference groups, respectively. As examples, for the Focal group we have

$$\hat{\delta}_{1,F}^E = -0.1638 + 0.6147 + 0 = 0.4509$$

$$\hat{\delta}_{2,F}^E = -0.1638 + 0.004629 + 0 = -0.15917$$

and for the Reference group we have

$$\hat{\delta}_{1,R}^E = -0.1638 + 1.567 + 0.6147 + -0.04515 = 1.97275$$

$$\hat{\delta}_{2,R}^E = -0.1638 + 1.567 + 0.004629 + 0.06549 = 1.473319.$$

22. The Laplace approximation yields the same result as the quadrature approach using a single quadrature point and is applicable to crossed random effects models (see https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_glimmix_<page number> where page number is “a0000001425.htm,” “a0000001426.htm,” “a0000001427.htm”; retrieved July 8, 2018). Because of computational efficiencies used with Laplace estimation, it is not as memory intensive as the quadrature approach. Thus, on occasion, one may receive an `out of memory/insufficient resources` error from SAS when using `method = quadrature`. Changing the estimation method to Laplace may allow one to perform the estimation.
23. If there is theory informing the selection of specific item characteristics and this framework describes the nature of the interrelationships, then use of an experimental design for testing the theory can allow one to establish causal relationships between this predictor and the criterion (e.g., item locations). As such, these models may be considered explanatory (i.e., explanatory IRT models). However, if there is no theoretical framework, then simply using item characteristics as predictors should not warrant the use of the term *explanatory*. This is particularly true for nonexperimental settings where it is impossible to isolate the variable(s) of interest to determine whether the observed relationships are spurious. Moreover, our item characteristics may simply be convenient proxies or abstractions (of the true cause(s)) useful for making predictions that describe performance on the item. In these cases, it may be prudent to not consider the item characteristics having an explanatory or *causal* interpretation. This principle of not conflating predictive modeling with explanatory modeling also applies to person-oriented characteristics.
24. As a simple example that exemplifies the need to, at times, treat items as random, consider that we wish to measure addition proficiency involving integers. Because

the number of integers is infinite, so are the number of item problems. Therefore, to create our instrument, we decide to limit ourselves to problems involving only two addends, each of which has one to four digits. Furthermore, we limit the test length to 50 items (i.e., 50 behavioral observations). We wish to infer from the examinee's performance their addition proficiency with *any* two addends that have one to four digits and not necessarily restrict ourselves to just those that appeared on our test. That is, we are not interested in the specific items on our test per se. As such, we consider our test to consist of a representative "random" sample from a specific population of addition items. Validity addresses whether our proficiency estimate inferences are appropriate. If we wish to couch this in a LLTM+ ϵ perspective, then our item characteristics could include whether the two addends have the same or opposite signs, the number of digits involved in the item problem, and in the case of opposite signs, whether the minuend is larger or smaller than the subtrahend. This idea can be applied to word problems as well as to instruments that are solely verbally focused; see Coleman (1964).

25. Our previous Rasch analyses treated items as fixed (constant). If we recalibrate the data treating items as fixed, the item location estimates of $\hat{\delta}_1^E = 2.25156$, $\hat{\delta}_2^E = 1.597285$, $\hat{\delta}_3^E = 1.18227$, etc. are almost identical to the values obtained from glimmix when we treat our items as fixed (except for differences in the fourth decimal).

```
> # Rasch analysis (fixed items)
> rasch=glmer(x~0+item+(1|person),family=binomial("logit"),data=stacked,
   control= optmzrinfo)
> summary(rasch)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + item + (1 | person)
Data: stacked
Control: optmzrinfo

AIC      BIC      logLik deviance df.resid
11180.6 11259.9  -5579.3  11158.6      9989

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.5615 -0.6676  0.2520  0.6544  4.1605

Random effects:
 Groups Name        Variance Std.Dev.
 person (Intercept) 0.5457    0.7387
 Number of obs: 10000, groups: person, 1000

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
itemi01   2.25156   0.10749  20.948 < 2e-16 ***
itemi02   1.59728   0.08718  18.322 < 2e-16 ***
itemi03   1.18227   0.07874  15.015 < 2e-16 ***

```

```

itemi04 0.52378    0.07105    7.372 1.68e-13 ***
itemi05 0.16357    0.06943    2.356  0.01848 *
itemi06 -0.22127   0.06960   -3.179  0.00148 **
itemi07 -0.34040   0.07004   -4.860  1.17e-06 ***
itemi08 -0.80080   0.07357  -10.885 < 2e-16 ***
itemi09 -1.46260   0.08419  -17.372 < 2e-16 ***
itemi10 -1.98590   0.09818  -20.227 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          item01 item02 item03 item04 item05 item06 item07 item08
item09
itemi02 0.062
itemi03 0.068  0.083
itemi04 0.073  0.090  0.099
itemi05 0.074  0.091  0.100  0.111
itemi06 0.072  0.089  0.099  0.110  0.113
itemi07 0.071  0.088  0.098  0.109  0.112  0.112
itemi08 0.067  0.083  0.092  0.103  0.107  0.107  0.107
itemi09 0.057  0.071  0.079  0.090  0.093  0.094  0.094  0.090
itemi10 0.048  0.060  0.067  0.076  0.079  0.080  0.080  0.078  0.070

> rasch_thetahat = coef(rasch)$person[, 1]                                # extracting people estimates
> people=matrix(rasch_thetahat)                                              # reformatting people
                                                                           estimates
> head(people,5)
 [,1]
[1,] -0.31261922
[2,] -0.31261922
[3,] -0.31261922
[4,] -0.31261922
[5,] -0.04068828

> tail(people,n=2)
 [,1]
[999,]  0.2311746
[1000,] -0.5867895

```

26. For comparison we present an LLTM analysis using `glmer`. Our η_s s are found in the Fixed Effects table ($\eta_1 = 1.39837$ and $\eta_2 = -1.07455$). The individual item location estimates are obtained by applying

$$\sum_{s=1}^S q_{js} \eta_s .$$

The estimated locations show a perfect correlation with those of a `glimmix` LLTM analysis ($\eta_1 = 1.3985$ and $\eta_2 = -1.07465$). With respect to `eRm` (Appendix E), we have $\eta_1 = 1.129$ and $\eta_2 = -1.597$. These values differ from those of `glmer` and `glimmix` because they are on a different scale. However, if we transform our `glmer` and `glimmix` η s to be on the same scale as those of `eRm`, then the corresponding η s are equal and the correlations among the estimated item locations for `eRm`, `glmer`,

and `glimmix` are 1.0. Moreover, in terms of AIC, BIC, and ΔG^2 (see `anova(LLTM, LLTMe)`), we have better model-data fit with the LLTM+ ε model than with the LLTM.

```
> # LLTM analysis
> LLTM=glmer(x~0+q1+ q2 +(1|person),family=binomial("logit"),data=stacked,
   control= optmzrinfo)

> summary(LLTM)
Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: x ~ 0 + q1 + q2 + (1 | person)
Data: stacked
Control: optmzrinfo

AIC      BIC      logLik deviance df.resid
12058.2 12079.8 -6026.1 12052.2     9997

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.8146 -0.6557  0.3553  0.7579  2.6501

Random effects:
Groups Name        Variance Std.Dev.
person (Intercept) 0.3944    0.628
Number of obs: 10000, groups: person, 1000

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
q1     1.39837   0.03807   36.73 <2e-16 ***
q2    -1.07455   0.03502  -30.68 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      q1
q2 -0.482

> anova(rasch,LLTM)
Data: stacked
Models:
LLTM: x ~ 0 + q1 + q2 + (1 | person)
rasch: x ~ 0 + item + (1 | person)
          Df  AIC  BIC  logLik deviance Chisq Chi Df
Pr(>Chisq)
LLTM       3 12058 12080 -6026.1     12052
rasch      11 11181 11260 -5579.3     11159 893.57      8 < 2.2e-16
***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> raschLLTM_thetahat = coef(LLTM)$person[, 1]
```

```
> peopleLLTM=matrix(raschLLTM _ thetahat)
> head(peopleLLTM,5)
[,1]
[1,] -0.20591020
[2,] -0.20591020
[3,] -0.20591020
[4,] -0.20591020
[5,]  0.01242182

> tail(peopleLLTM,n=2)
[,1]
[999,]  0.2305695
[1000,] -0.4256615

> anova(LLTM,LLTMe)  # Comparison w/ LLTM+e model
Data: stacked
Models:
LLTM: x ~ 0 + q1 + q2 + (1 | person)
LLTMe: x ~ 0 + (1 | item) + q1 + q2 + (1 | person)
      Df    AIC    BIC   logLik deviance Chisq Chi Df
Pr(>Chisq)
LLTM       3 12058 12080 -6026.1     12052
LLTMe      4 11226 11255 -5609.0     11218 834.23      1 < 2.2e-16
***  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

- Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.
- Ackerman, T. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement*, 18, 257–275.
- Ackerman, T. (1996). Developments in multidimensional item response theory. *Applied Psychological Measurement*, 20, 309–310.
- Ackerman, T. A. (1986, April). Use of the graded response IRT model to assess the reliability of direct and indirect measures of writing assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13, 113–127.
- Adams, R. J., & Khoo, S.-T. (1996). Quest: The Interactive Test Analysis System [Computer software]. Melbourne: Australian Council for Educational Research.
- Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory and practice* (Vol. 3, pp. 143–166). Norwood, NJ: Ablex.
- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1–23.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47–76.
- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Agresti, A. (1996). An introduction to categorical data analysis. New York: Wiley.
- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transaction Automatic Control*, 19, 716–723.
- Albano, A. D. (2016). equate: An R package for observed-score linking and \equating. *Journal of Statistical Software*, 74 (8), 1–36. Retrieved from <https://www.jstatsoft.org/article/view/v074i08>
- Albano, A. D. (2018). equate: Observed-score linking and equating [Computer software]. R package version 2.07. <https://CRAN.R-project.org/package=equate>
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: Author.
- Anastasi, A. (1983). Traits, states, and situations. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 345–356). Hillsdale, NJ: Erlbaum.
- Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimation. *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

- Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, 34, 42–50.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123–140.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–81.
- Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, 55, 1–25. Retrieved from <http://www.jstatsoft.org/v55/i06/>.
- Andersson, B., Bränberg, K., & Wiberg, M. (2020). *kequate: The Kernel Method of Test Equating* [Computer software]. R package version 1.63. <https://CRAN.R-project.org/package=kequate>
- Andrich, D. (1978a). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2, 449–460.
- Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Andrich, D. (1978c). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2, 581–594.
- Andrich, D. (1988). *Rasch models for measurement*. (Sage University Paper series on Quantitative Applications in the Social Sciences 07-068). Beverly Hills, CA: Sage.
- Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, 34, 8–14.
- Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.
- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Erlbaum.
- Assessment Systems Corporation. (1997). XCALIBRE [Computer software]. St. Paul, MN: Author.
- Bacci, S., & Caviezel, V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistics*, 38, 2775–2791.
- Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, 8, 261–271.
- Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement*, 14, 139–150.
- Baker, F. B. (1991). Comparison of minimum logit chi-square and Bayesian item parameter estimation. *British Journal of Mathematical and Statistical Psychology*, 44, 299–313.
- Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Dekker.
- Baker, F. B. (1993a). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, 17, 201–210.
- Baker, F. B. (1993b). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement*, 17, 20.
- Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. *Applied Psychological Measurement*, 20, 45–57.
- Baker, F. B. (1997). Empirical sampling distributions of equating coefficients for graded and nominal response instruments. *Applied Psychological Measurement*, 21, 157–172.
- Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147–162.
- Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement*, 15, 78.
- Baker, F. B., Cohen, A. S., & Baarmish, B. R. (1988). Item characteristics of tests constructed by linear programming. *Applied Psychological Measurement*, 12, 189–199.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.
- Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Cham, Switzerland: Springer International Publishing.
- Baker, F. B., & Subkoviak, M. J. (1981). Analysis of test results via log-linear models. *Applied Psychological Measurement*, 4, 503–515.
- Bandalos, D. L. (2018). *Measurement theory and applications for social sciences*. New York: Guilford Press.

- Barton, M. A., & Lord, F. M. (1981, July). *An upper asymptote for the three-parameter logistic item-response model* (Research Report No. 81-20). Princeton, NJ: Educational Testing Services.
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Battauz, M. (2013). IRT Test Equating in Complex Linkage Plans. *Psychometrika*, 78, 464–480.
- Battauz, M. (2015). equateIRT: An R Package for IRT Test Equating. *Journal of Statistical Software*, 68(7), 1–22.
- Battauz, M. (2018). equateIRT: IRT equating methods [Computer software]. R package version 2.1.0. <https://CRAN.R-project.org/package=equateIRT>
- Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement*, 1, 509–521.
- Bejar, I. I. (1986, June). *A psychometric analysis of a three-dimensional spatial task* (Research Report No. 86-19). Princeton, NJ: Educational Testing Services.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, 72, 200–223.
- Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39, 357–365.
- Berkson, J. (1955). Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association*, 50, 120–162.
- Bianconcini, S., & Cagnone, S. (2010). A multilevel latent variable model for multidimensional longitudinal data. In F. Palumbo, C. N. Lauro, & M. Greenacre (Eds.), *Data Analysis and Classification* (pp. 329–336). Berlin: Springer-Verlag.
- Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Bliese, P. (2016). Multilevel modeling in R (2.6). A brief introduction to R, the multilevel package and the nlme package. Retrieved from https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–49). New York: Springer.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179–198.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26, 381–409.
- Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics*, 30, 213–225.
- Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, 52, 465–484.
- Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier &

- C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). New York: Springer.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Bridgman, P. W. (1928). *The logic of modern physics*. New York: Macmillan.
- Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika*, 48, 269–291.
- Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2, 155–192.
- Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, 4, 379–426.
- Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 1, 287–316.
- Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., Thissen, D., & du Toit, S. H. C. (2020). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics*, 13, 227–241.
- Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, 16, 129–147.
- Camilli, G. (1994). Origin of the scaling constant $D = 1.7$ in item response theory. *Journal of Educational and Behavioral Statistics*, 19, 293–295.
- Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burkett. *Journal of Educational Measurement*, 36, 73–78.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Carlson, J. E. (1987, September). *Multidimensional item response theory estimation: A computer program* (Research Report No. ONR87-2). Iowa City, IA: American College Testing Program.
- Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 10, 1–19.
- Casabianca, J. M., Junker, B. W., Nieto, R., & Bond, M. A. (2017). A hierarchical rater model for longitudinal data. *Multivariate Behavioral Research*, 52, 576–592.
- Center for AIDS Prevention Studies. (2003). The Voluntary HIV Counseling and Testing Efficacy Study. www.caps.ucsf.edu/projects/c&tindex.html
- Cervantes, V. H. (2017a). DFIT: Differential functioning of items and tests [Computer software]. R package version 1.0.3. <https://CRAN.R-project.org/package=DFIT>
- Cervantes, V. H. (2017b). DFIT: An R package for Raju's differential functioning of items and tests framework. *Journal of Statistical Software*, 76(5), 1–24.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement*, 52, 200–222.
- Chalmers, R. P. (2017). Reference manual for package "mirt." Retrieved from <https://cran.r-project.org/web/packages/mirt/mirt.pdf>
- Chalmers, R. P. (2019). mirt: Multidimensional item response theory [Computer software]. R package version 1.31. <https://CRAN.R-project.org/package=mirt>
- Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2017). MLwiN Version 3.00. Centre for Multilevel Modelling, University of Bristol.

- Chen, H. (2012). A comparison between linear IRT observed-score equating and Levine observed-score equating under the generalized kernel equating framework. *Journal of Educational Measurement*, 49, 269–284.
- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
- Cheong, Y. K., & Kamata, A. (2013). Centering, scale indeterminacy, and differential item functioning detection in hierarchical generalized linear and generalized linear mixed models. *Applied Measurement in Education*, 26, 233–252.
- Childs, R. A., & Chen, W.-H. (1999). Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement*, 23, 371–379.
- Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, 35, 336–370.
- Cho, S.-J., Gilbert, J. K., & Goodwin, A. (2013). Explanatory multidimensional multilevel random item response model: An application to simultaneous investigation of word and person contributions to multidimensional lexical representations. *Psychometrika*, 78, 830–855.
- Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement*, 1, 114–142.
- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). *lordif*: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39(8), 1–30. URL <http://www.jstatsoft.org/v39/i08>
- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016a). *lordif*: Logistic ordinal regression differential item functioning using IRT [Computer software]. R package version 0.3-3. <https://CRAN.R-project.org/package=lordif>
- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016b). Reference manual for package “*lordif*.” Retrieved from <https://cran.r-project.org/web/packages/lordif/lordif.pdf>
- Chon, K. H., Lee, W., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, 47, 318–338.
- Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen’s Q_3 : Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, 41, 178–194.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 5–32.
- Cid, J. A., & von Davier, A. A. (2015). Examining potential boundary bias effects in Kernel smoothing on equating: An introduction for the adaptive and Epanechnikov kernels. *Applied Psychological Measurement*, 39, 208–222.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359.
- Cleary, T. A., Linn, R. L., & Rock, D. A. (1968). An exploratory study of programmed tests. *Educational and Psychological Measurement*, 28, 345–360.
- Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, 14, 219–226.
- Coombs, C. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, 145–158.
- Coombs, C. (1974). *A theory of psychological scaling*. In G. M. Maranell (Ed.), *Scaling: A sourcebook for behavioral scientists* (pp. 275–280). Chicago: Aldine. (Reprinted from Engineering Research Bulletin No. 34, University of Michigan, 1952.)
- Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London: Chapman and Hall.
- Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, 48, 129–141.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

- Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, 20, 405–416.
- Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.
- Dayton, C. M., & Scheers, N. J. (1997). Latent class analysis of survey data dealing with academic dishonesty. In J. Rost & R. L. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 172–180). Munich: Waxman Verlag.
- de Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, 16, 327–343.
- de Ayala, R. J. (1994). The influence of dimensionality on the graded response model. *Applied Psychological Measurement*, 18, 155–170.
- de Ayala, R. J. (2006). Estimating person locations from partial credit data containing missing responses. *Journal of Applied Measurement*, 4, 278–291.
- de Ayala, R. J., Dodd, B. G., & Koch, W. R. (1990). A computerized simulation of a flexilevel test and its comparison with a Bayesian computerized adaptive test. *Journal of Educational Measurement*, 27, 227–239.
- de Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2003). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2, 243–276.
- de Ayala, R. J., Plake, B., & Impara, J. (2001). The effect of omitted responses on ability estimation in IRT. *Journal of Educational Measurement*, 38, 213–234.
- de Ayala, R. J., & Sava-Bolesti, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, 23, 3–19.
- de Ayala, R. J., Schafer, W. D., & Sava-Bolesti, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British Journal of Mathematical and Statistical Psychology*, 48, 385–405.
- de Ayala, R. J., Smith, B., & Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Applied Psychological Measurement*, 42, 155–168.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533–559.
- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39(12), 1–28.
- De Champlain, A. F., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education*, 11, 231–253.
- De Champlain, A. F., & Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement*, 57, 174–178.
- de Gruijter, D. N. M. (1984). A comment on “Some standard errors in item response theory.” *Psychometrika*, 49, 269–272.
- de Gruijter, D. N. M. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement*, 27, 285–288.
- de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, 11, 57–85.
- DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, 27, 275–288.
- DeMars, C. E. (2007). “Guessing” parameter estimates for multidimensional item response theory models. *Educational and Psychological Measurement*, 67, 433–446.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Diamond, J. (2004). *Nutritional Literacy Scale*. Paper 1. Retrieved from <https://jdc.jefferson.edu/nls/1>
- Dillman, D. A. (2000). *Mail and Internet surveys*. New York: Wiley.
- Dillman, D. A., Eltinge, J. L., Groves, R. M., & Little, R. J. A. (2002). Survey nonresponse in design,

- data collection, and analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 3–26). New York: Wiley.
- Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, 1, 581–592.
- Divgi, D. R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, 44, 169–172.
- Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, 9, 413–415.
- Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 23, 283–298.
- Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models (Doctoral dissertation, University of Texas at Austin, 1984). Dissertation Abstracts International, 45, 2074A.
- Dodd, B. G. (1987, April). *Computerized adaptive testing with the rating scale model*. Paper presented at the Fourth International Objective Measurement Workshop, Chicago.
- Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, 14, 355–366.
- Dodd, B. G., & de Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. R. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 301–317). Norwood, NJ: Ablex.
- Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement*, 11, 371–384.
- Dodd, B. G., Koch, W. R., & de Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement*, 13, 129–144.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.
- Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter model. *Applied Psychological Measurement*, 13, 77–90.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59–79.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143–165.
- Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Hillsdale, NJ: Erlbaum.
- Dunn-Rankin, P., Knezeck, G. A., Wallace, S., & Zhang, S. (2004). *Scaling methods* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Duong, M. Q., & von Davier, A. A. (2008, March). *Kernel equating with observed mixture distribution in a single-group design*. Paper presented at the Annual Meeting of the National Council on Measurement in Education; New York.
- Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 57. New York: Chapman and Hall.
- Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175–186.
- Embretson, S. E. (1985). *Test design*. New York: Academic Press.
- Embretson, S. E. (1993). Learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Hillsdale, NJ: Erlbaum.
- Embretson, S. E. (1996). Cognitive design principles and the successful performer: A study on spatial ability. *Journal of Educational Measurement*, 33, 29–40.
- Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer.

- Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, 50, 328–344.
- Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, 55, 112–131.
- Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175–193.
- Emons, W., Sijtsma, K., & Meijer, R. (2004). Testing hypotheses about the person–response function in person–fit analysis. *Multivariate Behavioral Research*, 39, 1–35.
- Enders, C. K. (2001). A primer on maximum likelihood algorithms for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 128–141.
- Enders, C. K. (2003). Using the EM algorithm to estimate coefficient alpha for scales with item level missing data. *Psychological Methods*, 8, 322–337.
- Engelhard, G. (1990, April). Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Engelhard, G. (1994). Historical views of the concept of invariance in measurement theory. In M. R. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 73–99). Norwood, NJ: Ablex.
- Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, 6, 155–189.
- Engelhard, G. (2013). *Invariant measurement*. New York: Routledge.
- Enzmann, D. (2002). r_tetra: A SPSS-Macro [Computer software]. Hamburg, Germany: Institut für Kriminalwissenschaften, University of Hamburg. Available at www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html
- Falmagne, J. (1989). A latent trait theory via a stochastic learning theory for a knowledge space. *Psychometrika*, 54, 283–303.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council in Education and Macmillan.
- Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika*, 6, 323–329.
- Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item response. *Multivariate Behavioral Research*, 37, 521–542.
- Ferrando, P. J. (2004). Kernel-smoothing estimation of item characteristic functions for continuous personality items: An empirical comparison with the linear and the continuous-response models. *Applied Psychological Measurement*, 28, 95–109.
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, 45, 225–245.
- Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement*, 42, 149–169.
- Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, 46, 59–77.
- Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement*, 6, 397–416.
- Fischer, G. H., & Pendl, P. (1980). Individualized testing on the basis of the dichotomous Rasch model. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 171–188). New York: Wiley.
- Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, 59, 177–192.
- Fischer, G. H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 69–95). New York: Springer-Verlag.
- Fischer, G. H., & Seliger, E. (1997). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 323–346). New York: Springer.

- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1971a). On an absolute criterion for fitting frequency curves. In J. H. Bennett (Ed.), *The collected papers of R. A. Fisher* (Vol. 1, pp. 53–58). South Australia: University of Adelaide. (Original work published 1912)
- Fisher, R. A. (1971b). On the mathematical foundations of theoretical statistics. In J. H. Bennett (Ed.), *The collected papers of R. A. Fisher* (Vol. 1, pp. 310–368). South Australia: University of Adelaide. (Original work published 1921)
- Fischer, W. P. (1992). Reliability, separation, and strata statistics. *Rasch Measurement Transactions*, 6, 238.
- Flom, P. L., McMahon, J. M., & Pouget, E. R. (2007). Using proc nlmixed and proc glimmix to analyze dyadic data with a dichotomous dependent variable. *Proceedings of the SAS Global Forum 2007 Conference* (Paper 179-2007). Cary, NC: SAS Institute.
- Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement*, 5, 175–186.
- Fox, J.-P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, 15, 261–280.
- Fox, J. (2019). *polycor*: Polychoric and polyserial correlations [Computer Software]. R package version 0.7-10. <https://CRAN.R-project.org/package=polycor>
- Fraser, C. (1988). NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer software]. Armidale, New South Wales: Centre for Behavioural Studies, University of New England.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267–269.
- Fraser, C., & McDonald, R. P. (2003). NOHARM: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer software]. Welland, ON: Niagara College. Available at www.niagarac.on.ca/~cfraser/download
- Fraser, C., & McDonald, R. P. (2012). NOHARM 4: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer software]. Available at <https://cehs.unl.edu/edpsych/software-urls-and-other-interesting-sites/> or <https://noharm.software.informer.com/download>
- Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, 47, 432–457.
- Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.
- French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33, 315–332.
- French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, 47, 299–317.
- French, G. A., & Dodd, B. G. (1999). Parameter recovery for the rating scale model using PARSCALE. *Journal of Outcome Measurement*, 3, 176–199.
- Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology*, 14, 83–102.
- Frick, H., Leisch, F., Strobl, C., Wickelmaier, F., & Zeileis, A. (2020). *psychomix*: Psychometric Mixture Models [Computer software]. R package version 1.1-8. <http://CRAN.R-project.org/package=psychomix>
- Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, 48(7), 1–25.
- Gautschi, W. (1997). *Numerical analysis: An introduction*. Boston: Birkhäuser.
- Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8). Retrieved from <http://www.jtla.org>
- Gerald, C. F., & Wheatley, P. O. (1984). *Applied numerical analysis* (3rd ed.). Reading, MA: Addison-Wesley.
- Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the

- number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, 33, 157–179.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, 57, 423–436.
- Gierl, M. J., & Haladyna, T. M. (2013). Automatic item generation: Theory and practice. New York: Routledge.
- Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, 64, 273–294.
- Glas, C. A. W. (2007). Testing generalized Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 37–55). New York: Springer.
- Glas, C. A. W., & Dagohoy, V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, 72, 159–180.
- Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27, 87–106.
- Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, 68, 907–922.
- Glas, C. A. W., & Verhelst, N. D. (1995a). Testing the Rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer-Verlag.
- Glas, C. A. W., & Verhelst, N. D. (1995b). Tests of fit for polytomous Rasch models. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 325–352). New York: Springer-Verlag.
- Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, 73, 65–87.
- Goldstein, H. (1980). Dimensionality, bias independence, and measurement. *British Journal of Mathematical and Statistical Psychology*, 33, 234–246.
- Gonzalez, E. J., Smith, T. A., Sibberns, H., Adams, R., Dumais, J., Foy, P., et al. (1998). *User guide for the TIMSS international database: Final year of secondary school*. Chestnut Hill, MA: Boston College. Available at timss.bc.edu/
- Gonzalez, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating, *Journal of Statistical Software*, 59(7), 1–30. Retrieved from <http://www.jstatsoft.org/v59/i07>
- Gonzalez, J. (2020). SNSequate: Standard and nonstandard statistical models and methods for test equating [Computer software]. R package version 1.3.3 <https://CRAN.R-project.org/package=SNSequate>
- Goodman, L. A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis*. Cambridge, MA: Abt Books.
- Gourlay, N. (1951). Difficulty factors arising from the use of tetrachoric correlations in factor analysis. *British Journal of Psychology (Statistical section)*, 42, 65–76.
- Green, B. (1954). Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology* (pp. 355–369). Reading, MA: Addison-Wesley.
- Green, B. (1970). Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance* (pp. 184–197). New York: Harper & Row.
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347–360.
- Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, 12, 369–381.
- Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient Alpha. *Psychometrika*, 74, 155–167.
- Guilford, J. P. (1959). *Personality*. New York: McGraw-Hill.
- Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.
- Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum. (Originally published 1950)

- Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33, 205–233.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Guttman, L. (1950). The basis of scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.
- Haberman, S. J. (1978). Analysis of qualitative data: Vol. 1. *Introductory topics*. New York: Academic Press.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144–149.
- Häggstrom, J., & Wiberg, M. (2014). Optimal bandwidth selection in observed-score Kernel equating. *Journal of Educational Measurement*, 51, 201–211.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Hanson, B., & Zeng, L. (2004). ST: A computer program for IRT scale transformation [Computer software]. Iowa City, IA: ACT. Available at www.education.uiowa.edu/casma/IRTprograms.htm
- Harman, H. H. (1960). *Modern factor analysis*. Chicago: University of Chicago Press.
- Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101–125.
- Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, 15, 375–389.
- Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm. *Journal of Educational Statistics*, 13, 243–271.
- Harwell, M. R., Baker, F. B., & Zwarts, M. (1989). Correction: Harwell, Baker, and Zwarts, Vol. 13, No. 3. *Journal of Educational Statistics*, 14, 297.
- Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, 15, 279–291.
- Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139–164.
- Hattie, J. A., Krokowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, 20, 1–14.
- Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of graphic rating method. *Psychological Bulletin*, 18, 98–99.
- Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, and Winston.
- Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika*, 28, 211–218.
- Hetter, R. D., & Sympson, J. B. (1997). Item exposure in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.
- Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society, Series A*, 134, 486–501.
- Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, 64, 903–915.
- Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement*, 26, 337–349.
- Holman, R., & Glas, C. (2005, May). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58(Pt 1): 1–17.
- Holland, P. W. (1990a). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, 55, 5–18.
- Holland, P. W. (1990b). On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 557–601.
- Holland, P. W., & Hoskens (2003). Classical test theory as a first-order item response theory: application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68, 123–149.

- Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
- Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Technical Report No. 89-84). Princeton, NJ: Educational Testing Service.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed). New York: Wiley.
- Houts, C. R., & Cai, L. (2013). flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.
- Hu, L., & Bentler, P. M. (1998). Fit indexes in covariance structural equation modeling. *Psychological Methods*, 3, 424–453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Huang, H.-Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement*, 39, 362–372.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, 6, 249–260.
- Ingels, S. J., Scott, L. A., Rock, D. A., Pollack, J. M., & Rasinski, K. A. (1994). *National Education Longitudinal Study of 1988: First follow-up final technical report* (NCES 94–632). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.
- Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Hillsdale, NJ: Erlbaum.
- Jannarone, R. J., Yu, K. F., & Laughlin, J. E. (1990). Easy Bayes estimation for Rasch-type models. *Psychometrika*, 55, 449–460.
- Jansen, M. G. H. (1994). Parameters of the latent distribution in Rasch's Poisson Counts model. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 319–326). New York: Springer-Verlag.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer-Verlag.
- Jensema, C. J. (1974). The validity of Bayesian tailored testing. *Educational and Psychological Measurement*, 34, 757–766.
- Jensema, C. J. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, 36, 705–715.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, 6, 311–321.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329–349.
- Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, 35, 92–114.
- Jöreskog, K., & Sörbom, D. (1999). *PRELIS* (Version 2.30) [Computer software]. Mooresville, IN: Scientific Software.
- Kamata, A. (1998, April). *One-parameter hierarchical generalized linear logistic model: An application of HGLM to IRT*. Paper presented at the annual meeting of American Educational Research Association, San Diego.
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38, 79 – 93.
- Kaskowitz, G., & de Ayala, R. J. (2001). The effect of error in item parameter estimates error on the test response function method of linking. *Applied Psychological Measurement*, 25, 39–52.
- Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, 49, 223–245.
- Kendler, K. S., Karkowski, L. M., & Walsh, D. (1998). The structure of psychosis: Latent class analysis of probands from the Roscommon Family Study. *Archives of General Psychiatry*, 55, 492–499.

- Kim, D., de Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, 35, 447–471.
- Kim, S., & Kolen, M. J. (2003). POLYST: A computer program for polytomous IRT scale transformation [Computer software]. Iowa City, IA: University of Iowa. Program available at www.education.uiowa.edu/casma/IRTprograms.htm
- Kim, S.-H., & Cohen, A. S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. *Applied Psychological Measurement*, 16, 158.
- Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, 8, 291–312.
- Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22, 131–143.
- Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.
- Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, 43, 193–206.
- Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, 22, 197–206.
- Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14, 54–75.
- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26, 457–477.
- Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement*, 7, 15–32.
- Koch, W. R., & Reckase, M. D. (1979, September). *Problems in application of latent trait models to tailored testing* (Research Report No. 79-1). Columbia: University of Missouri, Tailored Testing Research Laboratory, Department of Educational Psychology.
- Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–275). New York: Plenum Press.
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer-Verlag.
- Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29, 285–307.
- Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33, 285–307.
- Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in psychometric research. *Educational and Psychological Measurement*, 69, 232–244.
- Kubinger, K. D., & Draxler, C. (2007). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 293–309). New York: Springer.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
- Larkin, K. C., & Weiss, D. J. (1975). *An empirical investigation of two-stage and pyramidal adaptive ability testing* (Research Report 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 61A, 273–287.
- Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh, Series A*, 62, 74–82.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.

- Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, 31, 121–134.
- Lei, P. W., Dunbar, S. B., & Kolen, M. (2004). A comparison of parametric and nonparametric approaches to item analysis for multiple-choice tests. *Educational and Psychological Measurement*, 64, 565–587.
- Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York: Academic Press.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161–176.
- Levine, M. V., Drasgow, F., & Stark, S. (2001). Program MODFIT [Computer software]. Urbana: University of Illinois, Measurement and Evaluation Laboratory, Department of Educational Psychology. Available at work.psych.uiuc.edu/irt/
- Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternate models for testlets. *Applied Psychological Measurement*, 30, 3–21.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5–55.
- Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, 75, 164–174.
- Linacre, J. M. (1994). *Many-faceted Rasch measurement*. Chicago: MESA Press, University of Chicago.
- Linacre, J. M. (1996). Cronbach Alpha, KR-20, True-score reliability or Rasch reliability? *Rasch Measurement Transactions*, 9, 455.
- Linacre, J. M. (1997). KR-20 / Cronbach Alpha or Rasch person reliability: Which tells the “truth”? *Rasch Measurement Transactions*, 11, 580–581.
- Linacre, J. M. (1999). Facets (Facets 3.83.0) [Computer software]. Chicago: MESA Press, University of Chicago.
- Linacre, J. M. (2001a). *A user’s guide to WINSTEPS/MINISTEPS*. Chicago: MESA Press.
- Linacre, J. M. (2001b). Facets Rasch measurement software [Computer Software]. Chicago, IL: Winsteps.com
- Linacre, J. M. (2002). What do Infit and Outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.
- Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions*, 17, 918.
- Linacre, J. M. (2004). Rasch model estimation: Further topics. In E. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 48–72). Maple Grove, MN: JAM Press.
- Linacre, J. M. (2005). *A user’s guide to WINSTEPS/MINISTEPS*. Chicago: MESA Press.
- Linacre, J. M., & Wright, B. D. (2001). *A user’s guide to BIGSTEPS*. Chicago: MESA Press.
- Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 96–107.
- Little, R., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Livingston, S. A. (1982). Estimation of the conditional standard error of measurement for stratified tests. *Journal of Educational Measurement*, 19, 135–138.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.
- Looken, E., & Rulison, K. L. (2010). Estimation of a 4-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63, 509–525.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.
- Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum’s three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989–1020.
- Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance* (pp. 139–183). New York: Harper & Row.
- Lord, F. M. (1971a). Robbins–Monro procedures for tailored testing. *Educational and Psychological Measurement*, 31, 3–31.
- Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement*, 8, 227–242.

- Lord, F. M. (1971c). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 31, 805–813.
- Lord, F. M. (1971d). A theoretical study of two-stage testing. *Psychometrika*, 36, 227–241.
- Lord, F. M. (1974a). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39, 247–264.
- Lord, F. M. (1974b). The relative efficiency of two tests as a function of ability level. *Psychometrika*, 39, 351–358.
- Lord, F. M. (1975). The “ability” scale in item characteristic curve theory. *Psychometrika*, 40, 205–217.
- Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1, 95–100.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.
- Lord, F. M. (1983a). Small N justifies the Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51–62). New York: Academic Press.
- Lord, F. M. (1983b). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48, 233–245.
- Lord, F. M. (1983c). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48, 477–482.
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23, 157–162.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179–193.
- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389–404.
- Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, 22, 224–236.
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 31, 19–26.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99.
- Magis, D. (2018). *difR*: Collection of methods to detect dichotomous differential item functioning (DIF) [Computer software]. R package version 5.0. <https://CRAN.R-project.org/package=difR>
- Magis, D., Béland, S., Tuerlinckz, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42, 847–862.
- Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, 37, 57–81.
- Maier, K. S. (2002). Modeling incomplete scaled questionnaire data with a partial credit hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, 27, 271–289.
- Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRM package for the application of IRT models in R. *Journal of Statistical Software*, 20(9), 1–20.
- Mair, P., Hatzinger, R., & Maier, M. J. (2018). *eRm*: Extended Rasch Modeling [Computer software]. R package version 0.16-2. <https://CRAN.R-project.org/package=eRm>
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Mao, X., von Davier, A. A., & Rupp, S. (2006). *Comparisons of the kernel equating method with the traditional equating methods on PRAXIS data (ETS RR-06-30)*. Princeton, NJ: Educational Testing Service.

- Marais, I. (2013). Local dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 111–130). London: Wiley.
- Maranell, G. M. (1974). *Scaling: A sourcebook for behavioral scientists*. Chicago: Aldine.
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139–160.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Masters, G. N. (1985). A comparison of latent-trait and latent-class analysis of Likert-type data. *Psychometrika*, 50, 69–82.
- Masters, G. N. (1988). Measurement models for ordered response categories. In R. L. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 11–29). New York: Plenum Press.
- Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, 49, 524–544.
- Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, 26, 51–71.
- Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713–732.
- McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.
- McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A. J., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, 43, 594–627.
- McCullagh, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.
- McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, No. 15.
- McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research*, 14, 21–38.
- McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100–117.
- McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.
- McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Erlbaum.
- McDonald, R. P., & Ahlswat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82–99.
- McDonald, R. P., & Mok, M. (1995). Goodness of fit in item response theory models. *Multivariate Behavioral Research*, 30, 23–40.
- McKinley, R. L., & Kingston, N. M. (1988, April). *Confirmatory analysis of test structure using multidimensional IRT*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9, 49–57.
- McKinley, R. L., & Reckase, M. D. (1983a, August). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report No. ONR83-2). Iowa City, IA: American College Testing Program.
- McKinley, R. L., & Reckase, M. D. (1983b). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation*, 15, 389–390.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107–135.
- Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105–118.
- Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19, 91–100.

- Mellenbergh, G. J., & Vijn, P. (1981). The Rasch model as a loglinear model. *Applied Psychological Measurement*, 5, 369–376.
- Microsoft Corporation. (2018). Microsoft Excel. Retrieved from <https://office.microsoft.com/excel>
- Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, 8, 271–288.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359–381.
- Mislevy, R. J. (1986a). Bayes modal estimation in item response models. *Psychometrika*, 51, 177–195.
- Mislevy, R. J. (1986b). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, 11, 3–31.
- Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, 42, 725–737.
- Mislevy, R. J., & Bock, R. D. (1985, April). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 189–202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.
- Mislevy, R. J., & Bock, R. D. (1997). BILOG 3: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.
- Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13, 57–75.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195–216.
- Mislevy, R. J., & Wu, P. (1988). *Inferring examinee ability when some item responses are missing* (RR 88-48-ONR). Princeton, NJ: Educational Testing Service.
- Mislevy, R. J., & Wu, P. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.
- Miyazaki, Y. (2005). Some links between classical and modern test theory via the two-level hierarchical generalized linear models. *Journal of Applied Measurement*, 6, 289–310.
- Miyazaki, Y., & Skaggs, G. (2008). Linking classical test theory and two-level hierarchical linear models. *Journal of Applied Measurement*, 9, 344–356.
- Moustaki, I., & O'Muircheartaigh, C. (2000). A one dimension latent trait model to infer attitude from nonresponse for nominal data. *Statistica*, 60, 259–276.
- Muckle, T. J., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement*, 46, 198–219.
- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59–71.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Muraki, E., & Bock, R. D. (2003). PARSCALE (Version 4.1) [Computer software]. Mooresville, IN: Scientific Software.
- Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement*, 9, 417–430.
- Muthén, B. O., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika*, 53, 563–578.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. Los Angeles: Muthén & Muthén.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.
- Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, 28, 99–117.
- Nash J. C., & Varadhan R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43(9), 1–14.
- National Opinion Research Center. (2003). General Social Science Survey. Available at www.icpsr.umich.edu:8080/GSS/homepage.htm
- Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, 19, 121–129.

- Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, 22, 53–69.
- Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1–32.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535–569.
- O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 191–242). Charlotte, NC: Information Age.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289–298.
- Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31, 200–219.
- Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, 37, 357–373.
- Panter, A. T., Swygert, K. A., Dahlstrom, W. G., & Tanaka, J. S. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment*, 68, 561–589.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computerized-based testing*. New York: Springer.
- Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education*, 16, 223–243.
- Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Modeling*, 30, 100–120.
- Patience, W. M. (1977). Description of components in tailored testing. *Behavior Research Methods and Instrumentation*, 9, 153–157.
- Patience, W. M., & Reckase, M. D. (1980, April). Effects of program parameters and item pool characteristics on the bias of a three-parameter tailored testing procedure. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.
- Patterson, D. G. (1922). The Scott Company graphic rating scale. *Journal of Personnel Research*, 1(8–9), 361–376.
- Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146–178.
- Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342–366.
- Paxton, P., Curran, P. J., Bollen, K., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8, 287–312.
- Pedhauzer, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt Brace.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.
- Pinheiro, J., Bates D., DebRoy S., Sarkar D., & R Core Team (2018). nlme: Linear and nonlinear mixed effects models [Computer software]. R package version 3.1-137. <https://CRAN.R-project.org/package=nlme>
- Preinerstorfer, D. (2016). mRm: An R package for conditional maximum likelihood estimation in mixed Rasch models [Computer software]. R package version 1.1.6. <https://CRAN.R-project.org/package=mRm>
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>

- Rabe-Hesketh, S., & Skrondal, A., (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1), 1–21.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197–207. (A correction is found in *Applied Psychological Measurement*, 15, 352.)
- Ramsay, J. O. (1989). A comparison of three simple test theory models. *Psychometrika*, 54, 487–499.
- Randall, J., Cheong, Y. F., & Engelhard, G. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, 71, 129–147.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321–333). Berkeley: University of California Press.
- Rasch, G. (1966, July). An individual-centered approach to item analysis with two categories of answers. In L. J. T. van der Kamp & C. A. J. Vlek (Eds.), *Psychological measurement theory*. Proceedings of the NUFFIC international summer session in science at “Het Oude Hof.” The Hague: Netherlands.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960)
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: SAGE.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). Hierarchical linear and nonlinear modeling [Computer software]. Lincolnwood, IL: Scientific Software.
- Reckase, M. D. (1977). Procedures for computerized testing. *Behavior Research Methods and Instrumentation*, 9, 208–212.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207–230.
- Reckase, M. D. (1980, April). *An application of tailored testing and sequential analysis to classification problems*. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Reckase, M. D. (1985). The difficulty of tests that measure more than one ability. *Applied Psychological Measurement*, 9, 401–412.
- Reckase, M. D. (1986, April). *The discriminating power of items that measure more than one dimension*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Reckase, M. D. (1989, Fall). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice*, 8(3), 11–15.
- Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.
- Reckase, M. D. (1997b). Past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25–36.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.
- Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25, 193–203.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15, 361–373.
- Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 135–146). New York: Academic Press.
- Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model–data fit in IRT. *Applied Psychological Measurement*, 14, 127–137.
- Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, 35, 543–568.

- Reise, S. P., & Due, N. G. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, 15, 217–226.
- Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, 14, 45–58.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27, 133–144.
- Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, 38, 549–562.
- Revelle, W. (2018). psych: Procedures for personality and psychological research [Computer software]. R package version 1.8.12. Northwestern University, Evanston, IL. <https://CRAN.R-project.org/package=psych> Version = 1.8.12
- Ricker, K. L., & von Davier, A. (2007). The impact of anchor test length on equating results in a non-equivalent groups design (RR-07-44). Princeton, NJ: Educational Testing Service.
- Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47, 361–372.
- Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic model. *Applied Psychological Measurement*, 26, 271–285.
- Roberts, J. K., & Herrington, R. (2005). Demonstration of software programs for estimating multilevel measurement model parameters. *Journal of Applied Measurement*, 6, 255–272.
- Robitzsch, A. (2018). sirt: Supplementary item response theory models [Computer software]. R package version 3.5-53. <https://CRAN.R-project.org/package=sirt>
- Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis modules [Computer software]. R package version 3.5-19. <https://CRAN.R-project.org/package=TAM>
- Rose, N., von Davier, M., Xu, X. (2010). Nonignorable missing data with item response theory (IRT) (Research Report ETS-10-11). Princeton, NJ: Educational Testing Service.
- Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49, 425–435.
- Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282.
- Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44, 75–92.
- Rost, J., & von Davier, M. (1992). MIRA: A PC-program for the mixed Rasch model [Computer software]. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel.
- Roth, R. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537–560.
- Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, 64, 588–599; Errata, 64, 991.
- Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66, 63–84.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.
- Samejima, F. (1973a). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, 38, 221–233.
- Samejima, F. (1973b). Homogeneous case of the continuous response model. *Psychometrika*, 38, 203–219.
- Samejima, F. (1974). A normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39, 111–121.
- Samejima, F. (1976). The graded response model of latent trait theory and tailored testing. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (U.S. Civil Service

- Commission, Personnel Research and Development Center, PS-75-6) (pp. 5–15). Washington, DC: U.S. Government Printing Office.
- Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report No. 79-4). Knoxville: University of Tennessee, Department of Psychology.
- Samejima, F. (1983). *A general model for the homogeneous case of the continuous response* (Research Report No. 83-3). Knoxville: University of Tennessee, Department of Psychology.
- Samejima, F. (1990). Predictions of reliability coefficients and standard errors of measurement using the test information function and its modifications (ONR/RR-90-2). Knoxville: University of Tennessee, Department of Psychology.
- Samejima, F. (1994). Some critical observations of the test information function as a measure of local accuracy in ability estimation. *Psychometrika*, 59, 307–329.
- Samejima, F. (2001). *Non-parametric on-line item calibration* (Law School Admission Council, 1999–2001, Final Report). Knoxville: University of Tennessee, Department of Psychology.
- Samejima, F. (2010). The general graded response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77–107). New York: Taylor & Francis.
- Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- SAS Institute. (2012). *SAS for Windows*: Version 9.4. Carey, NC: Author.
- Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. *Sociological Methodology*, 18, 271–307.
- Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.
- Seagall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331–354.
- Seliger, E., & Fischer, G. H. (1994). LRSMG, LRSM, LLTM, LPCM: Program description with applications to scale analysis and measuring change [Computer software]. Vienna, Austria: University of Vienna.
- Seong, T.-J. (1990a, April). Validity of using two numerical analysis techniques to estimate item and ability parameters via MMLE: Gauss–Hermite quadrature formula and Mislevy’s histogram solution. Paper presented at the meeting of the National Council of Measurement in Education, Boston.
- Seong, T.-J. (1990b). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, 14, 299–311.
- Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317–375.
- Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, 9, 93–128.
- Shojima, K. (2005). A noniterative item parameter solution in each EM cycle of the continuous response model. *Educational Technology Research*, 28, 11–22.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika*, 74, 107–120.
- Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational Statistics*, 25, 391–415.
- Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person fit research. *Psychometrika*, 66, 191–208.
- Silk, K. J., Sherry, J., Winn, B., Keesecker, N., Horodynski, M. A., & Sayir, A. (2008). Increasing nutrition literacy: Testing the effectiveness of print, web site, and game modalities. *Journal of Nutrition Education and Behavior*, 40, 3–10.
- Sinharay, S., Haberman, S., Holland, P. W., & Lewis, C. (2012). A note on the choice of an anchor test in equating (RR-12-14). Princeton, NJ: Educational Testing Service.
- Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests

- being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, 44, 249–275.
- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.
- Sireci, S. G., Wainer, H., & Thissen, D. J. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237–247.
- Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-Bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. *Applied Psychological Measurement*, 13, 391–402.
- Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series B*, 172 (Part 3), 659–687.
- Smith, E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 93–122). Maple Grove, MN: JAM Press.
- Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, 45, 433–444.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541–565.
- Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.
- Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2, 66–78.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331–342.
- Spada, N., & McGaw, B. (1985). Learning and cognitive processes. In S. E. Embretson (Ed.), *Test design* (pp. 169–194). New York: Academic Press.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 32, 583–639.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). WinBUGS [Computer software]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health. Available at www.mrc-bsu.cam.ac.uk/bugs/
- Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990, August). Comparison of two logistic multidimensional item response theory models (Research Report No. ONR90-8). Iowa City, IA: American College Testing Program.
- SPSS Incorporated (2019). SPSS 26.0 for Windows. Chicago: Author.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173–180.
- Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, 81, 332–342.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Stocking, M. L., Eignor, D., & Cook, L. (1988). Factors affecting the sample invariant properties of linear and curvilinear observed and true score equating procedures (RR-88-41). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23, 57–75.
- Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Dordrecht, The Netherlands: Kluwer.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, 16, 1–16.
- Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistics in IRT models. *Journal of Educational Measurement*, 37, 58–75.
- Stone, C. A., & Zhang, B. J. (2003). Assessing goodness of fit of item response theory models: A

- comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40, 331–352.
- Stone, M. H., Wright, B. D., & Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement*, 3, 308–322.
- Stouffer, S. A. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 3–45). Princeton, NJ: Princeton University Press.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589–617.
- Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293–325.
- Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice-Hall.
- Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement*, 21, 49–58.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics—Theory and Methods*, 7, 13–26.
- Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *Journal of Early Adolescence*, 37, 85–128.
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175–191.
- Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.
- Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349–364.
- Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51, 589–601.
- Swaminathan, H., & Rogers, J. (1990). Detecting DIF using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27, 53–75.
- Swenson, W. M., Pearson, J. S., & Osborne, D. (1973). *An MMPI source book: Basic item, scale, and pattern data on 50,000 medical patients*. Minneapolis: University of Minnesota Press.
- Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- SYSTAT (2017). *SYSTAT Version 13*. San Jose, CA: Author.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393–408.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354.
- Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal Psychology*, 48, 285–290.
- Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, 74(5), 1–27.
- Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2018). *PerFit*: Person fit [Computer software]. R package version 1.4.3. <https://CRAN.R-project.org/package=PerFit>
- Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika*, 50, 411–420.
- Thissen, D. J. (1976). Information in wrong responses to Raven's progressive matrices. *Journal of Educational Measurement*, 13, 201–214.

- Thissen, D. J. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175–186.
- Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software and manual]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory, University of North Carolina.
- Thissen, D. J., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *The handbook of item response theory, Volume 1: Models* (pp. 51–73). Boca Raton, FL: CRC Press.
- Thissen, D. J., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering and R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43–75). New York: Routledge.
- Thissen, D. J., Chen, W.-H., & Bock, R. D. (2003). MULTILOG (Version 7.0) [Computer software]. Mooresville, IN: Scientific Software.
- Thissen, D. J., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501–519.
- Thissen, D. J., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Thissen, D. J., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–65). New York: Springer.
- Thissen, D. J., Steinberg, L., & Fitzpatrick, A. R. (1989a). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26, 161–176.
- Thissen, D. J., Steinberg, L., & Mooney, J. A. (1989b). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26, 247–260.
- Thissen, D. J., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.
- Thissen, D. J., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.
- Thissen, D. J., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47, 397–412.
- Thomson, William (Lord Kelvin). (1891). *Popular lectures and addresses* (Vol. 1). New York: Macmillan.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433–451.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.
- Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, No. 1.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Torres Irribarra, D., & Freund, R. (2014). WrightMap: IRT item-person map with ConQuest integration [Computer software]. R package version 1.2.1. <https://CRAN.R-project.org/package=WrightMap>
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.
- Tryon, R. C. (1935). A theory of psychological components – An alternative to “mathematical factors”. *Psychological Review*, 42, 425–454.
- Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin*, 54, 229–249.
- Tryon, R. C. (1959). Domain sampling formulation of cluster and factor analysis. *Psychometrika*, 24, 113–135.
- Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1–13.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39–55.

- Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 34, 253–269.
- Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 14, 181–196.
- Vale, C. D., Albing, C., Foote-Lennox, L., & Foote-Lennox, T. (1982). *Specification of requirements and preliminary design* (RR ONR-ASC-82-01). St. Paul, MN: Assessment Systems.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://www.jstatsoft.org/v45/i03/>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2019). mice: Multivariate imputation by chained equations [Computer software]. R package version 3.6.0. <https://CRAN.R-project.org/package=mice>
- Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, 30, 443–464.
- Van den Noortgate, W., De Boeck, P., & Meuldert, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28, 369–386.
- van den Wollenberg, A. (1988). Testing a latent trait model. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 31–50). New York: Plenum Press.
- van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13, 267–298.
- van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195–211.
- van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 24, 398–412.
- van der Linden, W. J., & Boekkooi-Timmeringa, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, 54, 237–247.
- van der Linden, W. J., & Glas, C. A. W. (Eds.) (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, the Netherlands: Kluwer.
- van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. New York: Springer.
- van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29, 273–291.
- van Nispen, R. M. A., Knol, D. L., Neve, H. J., & van Rens, G. H. M. B. (2010). A multilevel item response theory model was investigated for longitudinal vision-related quality-of-life data, *Journal of Clinical Epidemiology*, 63, 321–330.
- Verhelst, N. D., & Glas, C. A. W. (1995). One-parameter logistic model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–237). New York: Springer-Verlag.
- Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). One-parameter logistic model (OPLM). Arnhem, the Netherlands: CITO.
- Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). New York: Springer-Verlag.
- Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York: Springer.
- Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213–239.
- Vermunt, J. K. (2007). Multilevel mixture item response theory models: An application in education testing. *Bulletin of the International Statistical Institute*, 56th Session, paper #1253, 1–4. ISI 2007: Lisboa, Portugal. Retrieved from <https://jeroenvermunt.nl>
- von Davier, M. (2001). WINMIRA 2001 [Computer software]. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel. Available at www.winmira.von-davier.de
- von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution Rasch models*. New York: Springer.
- von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data* (ETS RR-06-02). Princeton, NJ: Educational Testing Service.

- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating*. New York: Springer.
- Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 63–80). Hillsdale, NJ: Erlbaum.
- Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht, The Netherlands: Kluwer.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory*. New York: Cambridge University Press.
- Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing. *Journal of Educational Measurement*, 27, 1–14.
- Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 24, 185–201.
- Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, 45, 373–391.
- Walker, D. A. (1931). Answer pattern and score scatter in tests and examinations. *British Journal of Psychology*, 22, 73–86.
- Walker-Barnick, L. A. (1990). An investigation of factors affecting invariance of item parameter estimates in the partial credit model. Unpublished doctoral dissertation, University of Maryland, College Park.
- Wang, M. D. (1986, April). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the ONR Contractors Conference, Gatlinburg, TN.
- Wang, M. D. (1987, April). *Estimation of ability parameters from response data to items that are pre-calibrated with a unidimensional model*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Wang, T., & Zeng, L. (1998). Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement*, 22, 333–344.
- Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29, 126–149.
- Wang, X. B., Wainer, H., & Thissen, D. J. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education*, 8, 211–225.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450.
- Way, W. D. (1998). Protecting the integrity of computerized adaptive testing item pools. *Educational Measurement: Issues and Practice*, 17(4), 17–27.
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, 12, 239–252.
- Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, 35 (12), 1–33. Retrieved from <http://www.jstatsoft.org/v35/i12>
- Weeks, J. P. (2017). plink: IRT Separate Calibration Linking Methods [Computer software]. R package version 1.5.1. <https://CRAN.R-project.org/package=plink>
- Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. (NTIS No. AD 768376).
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.
- Weiss, D. J. (1983). Introduction. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 1–8). New York: Academic Press.
- Weitzman, R. A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement*, 56, 779–790.

- Whately (Embretson), S. E. (1977). Models, meanings and misunderstandings: Some issues in applying Rasch's theory. *Journal of Educational Measurement*, 14, 227–235.
- Whately (Embretson), S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479–494.
- Wickham, H. (2007). Reshaping data with the `reshape` package. *Journal of Statistical Software*, 21(12), 1–20. URL <http://www.jstatsoft.org/v21/i12>
- Wickham, H. (2017). `tidyverse`: Easily install and load the “tidyverse” [Computer software]. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H. (2019). `tidyrr`: Easily tidy data with “`spread()`” and “`gather()`” functions [Computer software]. R package version 0.8.3. <https://CRAN.R-project.org/package=tidyr>
- Willse, J. (2011). Mixture Rasch models with joint maximum likelihood estimation. *Educational and Psychological Measurement*, 71, 5–19.
- Willse, J. (2014). `mixRasch`: Mixture Rasch models with JMML [Computer software]. R package version 1.1. <https://CRAN.R-project.org/package=mixRasch>
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
- Wilson, M., & Adams, R. J. (1993). Marginal maximum likelihood estimation for the partial ordered model. *Journal of Educational Statistics*, 18, 69–90.
- Wilson, M., & Masters, G. N. (1993). The partial credit model and null categories. *Psychometrika*, 58, 87–99.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide [Computer software]. Princeton, NJ: Educational Testing Service.
- Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, 26, 339–352.
- Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). TESTFACT (Version 4.0) [Computer software]. Mooresville, IN: Scientific Software.
- Wood, R. L. (1973). Response-contingent testing. *Review of Educational Research*, 43, 529–544.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.
- Wright, B. D. (1977a). Misunderstanding the Rasch model. *Journal of Educational Measurement*, 14, 219–226.
- Wright, B. D. (1977b). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97–116.
- Wright, B. D. (1980). Afterword. In G. Rasch, *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, 3, 281–288.
- Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9, 472.
- Wright, B. D., Congdon, R., & Shultz, M. (1988). MSTEPs partial credit analysis [Computer software]. Chicago: University of Chicago, MESA Psychometric Laboratory.
- Wright, B. D., & Douglas, D. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1, 281–295.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). ConQuest: Multi-aspect test software [Computer software]. Camberwell: Australian Council for Educational Research.
- Yamamoto, K. (1989). A HYBRID model of IRT and latent class models (RR-89-41). Princeton, NJ: Educational Testing Service.
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30, 469–492.

- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399–410.
- Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23, 299–326.
- Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52, 275–291.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.
- Yen, W. M., Burkett, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika*, 56, 39–54.
- Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2004). RAGE -RGEQUATE [Computer software]. Iowa City: University of Iowa. Program available at www.education.uiowa.edu/casma/EquatingLinkingPrograms.htm
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.
- Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3.0) [Computer software]. Mooresville, IN: Scientific Software.
- Zoellner, J., You, W., Connell, C., Smith-Ray, R. L., Allen, K., Tucker, K. L., Davy, B. M., & Estabrooks, P. A. (2011). Health literacy is associated with healthy eating index scores and sugar-sweetened beverage intake: Findings from the Rural Lower Mississippi Delta. *Journal of the American Dietetic Association*, 111, 1012–1020.
- Zopluguoglu, C. (2012). EstCRM: An R package for Samejima's continuous IRT model. *Applied Psychological Measurement*, 36, 149–150.
- Zopluguoglu, C. (2013). A comparison of two estimation algorithms for Samejima's continuous IRT model. *Behavior Research Methods*, 45, 54–64.
- Zopluguoglu, C. (2015). EstCRM: Calibrating parameters for the Samejima's continuous IRT model [Computer software]. R Package version 1.4. <https://CRAN.R-project.org/package=EstCRM>
- Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.
- Zwenderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, 56, 589–600.
- Zwenderman, A. H., & van der Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, 14, 73–81.

Author Index

- Ackerman, T., 76, 407, 481
Ackerman, T. A., 333, 392, 411, 439
Adams, R. J., 44, 78, 244, 525, 527, 557, 587, 588
Agresti, A., 359, 384, 485, 488
Ahlawat, K. S., 48
Aitkin, M., 87, 88, 92, 93, 405, 421, 436, 440
Akaike, H., 152
Albano, A. D., 444
Al-Dosary, I. M., 453
Al-Karni, A., 450, 453
Anastasi, A., 10
Andersen, E. B., 44, 85, 86, 267
Andrich, D., 12, 20, 37, 168, 267, 269, 272, 297, 307
Angoff, W. H., 444, 445, 446, 479
Ansley, T. N., 439
Averill, M., 174

Baarmish, B. R., 35
Bacci, S., 588
Baker, F. B., 35, 40, 43, 44, 45, 88, 92, 93, 126, 140, 183, 185, 329, 364, 450, 453, 455, 475
Bandalos, D. L., 47, 75
Barton, M. A., 44, 227
Bates, D., 531, 541
Battauz, M., 450
Bejar, I. I., 347, 349, 351
Béland, S., 151, 213, 520
Bentler, P. M., 248
Beretvas, S. N., 525
Berger, M. P. F., 48
Bernstein, I. H., 3
Best, N., 227
Best, N. G., 159
Bianconcini, S., 578
Birnbaum, A., 29, 31, 41, 163

Bliese, P., 531
Bock, R. D., 45, 78, 87, 88, 92, 93, 95, 96, 97, 98, 102, 105, 141, 172, 173, 248, 252, 293, 301, 360, 361, 365, 385, 405, 421, 436, 440
Boekkooi-Timmeringa, E., 35
Bolker, B. M., 531, 541
Bolt, D. M., 229, 369, 370, 438
Bond, M. A., 530
Bonifay, W., 438
Bradlow, E. T., 189
Brennan, R. L., 40, 41, 444, 445, 446, 476
Bridgman, P. W., 1, 2
Brown, C. H., 222
Brown, J. S., 357
Brown, R. L., 222
Browne, M. W., 376
Browne, W. J., 531
Bryk, A. S., 531, 588
Burket, G. R., 186
Burton, R. R., 357
Bush, M. J., 58

Cagnone, S., 578
Cai, L., 45, 248, 253, 301, 365, 438
Cameron, B., 531
Camilli, G., 174, 475, 479, 482, 485
Campbell, D. T., 76
Carlin, B. P., 159
Carlson, J. E., 392, 405, 411
Carstensen, C. H., 225
Casabianca, J. M., 530
Case, S. M., 547
Caviezel, V., 588
Cervantes, V. H., 151, 520
Chalmers, R. P., 45, 98, 370, 465
Charlton, C., 531

- Chen, W.H., 141, 188, 190, 191, 252, 293
 Cheong, Y. F., 525, 531
 Cheong, Y. K., 545, 546
 Chien, Y., 444
 Cho, S. J., 525, 547, 579
 Choi, S. W., 293, 515
 Chon, K. H., 159, 176
 Christensen, K. B., 191
 Clauser, B. E., 547
 Cohen, A. S., 35, 222, 369, 446, 447, 482, 520,
 547, 579
 Cohen, J., 263
 Coleman, E. B., 593
 Congdon, R., 292
 Congdon, R. T., 531
 Cook, K. F., 293
 Cook, L., 223
 Coombs, C., 3, 4, 20
 Cox, D. R., 524
 Craig, R. D., 491
 Cramér, 41
 Crane, P. K., 515
 Cressie, N., 226, 229
 Cronbach, L. J., 84, 569
 Cui, Z., 444
- Dagohoy, V. T., 98
 Dahlstrom, W. G., 48
 Davey, T. C., 392, 454
 Dayton, C. M., 8
 de Ayala, R. J., 97, 189, 223, 295, 296, 366,
 368, 369, 392, 450
 De Boeck, P., 5, 151, 222, 520, 527, 529, 544,
 547, 567, 568, 569, 574, 588
 DebRoy S., 531
 De Champlain, A. F., 48
 de Gruijter, D. N. M., 87, 185
 de Leeuw, J., 588
 DeMars, C. E., 368, 369
 Dempster, A. P., 92
 Diamond, J., 567
 Dillman, D. A., 46, 222
 Dinero, T. E., 224
 Divgi, D. R., 477
 Dodd, B. G., 242, 293, 295, 296, 307, 329, 333
 Dorans, N. J., 482, 486
 Douglas, D. A., 83, 84
 Drasgow, F., 93, 141, 162, 171, 209, 212
 Draxler, C., 226
 Du, Z., 189
 du Toit, M., 531
 du Toit, S. H. C., 45
 Dunbar, S. B., 159, 176
- Dunn-Rankin, P., 3
 Efron, B., 421
 Eignor, D., 223
 Eltinge, J. L., 222
 Embretson, S. E., 5, 436, 567, 569
 Enders, C. K., 222
 Engelhard, G., 8, 74, 525
- Falcón, J. C. S., 189
 Falmagne, J., 5
 Featherman, C., 547
 Feldt, L. S., 40
 Ferdous, A. A., 189
 Ferguson, G. A., 48
 Ferrando, P. J., 351
 Finch, H., 48, 223
 Finch, W. H., 482, 547
 Fischer, G., 530
 Fischer, G. H., 5, 186, 436
 Fisher, R. A., 30, 39, 40
 Fiske, D. W., 76
 Fitzpatrick, A. R., 367
 Flom, P. L., 589
 Formann, A. K., 5
 Forsyth, R., 224
 Forsyth, R. A., 439
 Fox, J.P., 525, 530, 588
 Fraser, C., 48, 52, 405, 406
 French, A. W., 490
 French, B. F., 482, 547
 French, G. A., 293
 Freund, R., 74
 Freyd, M., 344
 Fu, J., 229, 438
- Gessaroli, M. E., 48
 Gibbons, L. E., 515
 Gibbons, R. D., 438
 Gierl, M. J., 490, 569
 Gifford, J. A., 87, 95, 185, 229
 Gilbert, J. K., 525
 Gilmer, J., 224
 Glas, C. A. W., 44, 98, 171, 189, 223, 225
 Gleser, G., 569
 Goegebeur, Y., 222
 Goldstein, H., 22
 Goldstein, J., 589
 Gonzalez, E. J., 370
 Gonzalez, J., 444, 449, 450, 463
 Goodwin, A., 525
 Green, B., 17

- Green, B. F., 172, 173
 Groothuis-Oudshoorn, K., 223
 Groves, R. M., 222
 Guilford, J. P., 10
 Gulliksen, H., 3, 8
 Gustafsson, J. E., 225, 228
 Guttman, L., 3, 38, 39, 77, 84
- Haberman, S., 446
 Haberman, S. J., 199
 Habing, B., 48
 Haebara, T., 450, 476
 Haenszel, W., 483
 Haertel, E., 224
 Haladyna, T. M., 569
 Hambleton, R. K., 114
 Hanson, B., 450
 Hanson, B. A., 41, 444
 Harwell, M. R., 92, 93, 141
 Hattie, J. A., 39, 48
 Hatzinger, R., 44
 Hayes, M. H. S., 344
 Hays, W. L., 87
 Healy, M., 531
 Hedeker, D. R., 438
 Hemker, B. T., 364
 Henry, 541
 Herrington, R., 536
 Hidalgo, M. D., 482
 Hirsch, T. M., 410, 440, 454
 Holland, P. W., 8, 226, 229, 444, 446, 479,
 482, 483, 485, 486
 Holman, 223
 Hong, S., 45
 Hoover, H. D., 445, 449
 Hornke, L. F., 525
 Horton, M., 191
 Hoskens, 8
 Hosmer, D. W., 487, 488
 Houts, C. R., 45
 Hu, L., 248
 Huang, H.Y., 525, 588
 Hulin, C. L., 171, 229
 Humphreys, L. G., 172
- Impara, J., 223
 Ingels, S. J., 185
 Ironson, G. H., 491
- Jannarone, R. J., 87
 Janosky, J. E., 93, 141
 Jansen, M. G. H., 22, 37
- Janssen, R., 568, 569
 Jarjoura, D., 446
 Jensen, A. R., 479
 Jensen, H. E., 447, 449
 Jiao, H., 525, 578
 Jodoin, M. G., 490
 Joe, H., 121
 Johnson, T. R., 370
 Junker, B. W., 126, 530
- Kamata, A., 525, 527, 545, 546, 557, 574, 578,
 588
 Karabatsos, G., 530
 Karkowski, L. M., 8
 Kaskowitz, G., 450
 Kendler, K. S., 8
 Khoo, S.-T., 44
 Kiefer, T., 44
 Kiely, G. L., 188
 Kim, D., 189
 Kim, S., 450
 Kim, S. H., 40, 43, 44, 93, 126, 140, 185, 446,
 447, 482, 520
 Kingston, N. M., 405, 569
 Klein, L. W., 446
 Klein Entink, R. H., 525
 Knezek, G. A., 3
 Knol, D. L., 48, 588
 Koch, W. R., 295, 333, 447
 Kok, F., 76
 Kolen, M. J., 41, 444, 445, 446, 450, 476, 556
 Kreft, I., 588
 Krokowski, K., 39
 Kubinger, K. D., 226, 567
 Kuhn, J. T., 525
 Kuhn, T. S., 5
- Laird, N. M., 92
 Laughlin, J. E., 87
 Lazarsfeld, P. F., 10, 11, 14, 17
 Lee, K., 454
 Lee, W., 159, 176
 Lee, Y.S., 369
 Lemeshow, S., 487, 488
 Levine, M. V., 162, 209, 212
 Lewis, C., 188, 446
 Li, Y., 229, 438
 Lieberman, M., 87
 Likert, R., 267
 Lim, R. G., 171
 Linacre, J. M., 44, 56, 58, 59, 62, 65, 74, 122,
 277, 278, 281, 283, 306, 310
 Linacre, M., 40

- Linn, R. L., 172
 Lissak, R. I., 171
 Little, R., 221
 Little, R. J. A., 222
 Livingston, S. A., 40, 41, 443, 444
 López-Pina, 482
 Lord, F. M., 6, 7, 17, 22, 29, 38, 41, 43, 44, 79,
 87, 88, 95, 113, 114, 140, 163, 165, 169,
 170, 182, 183, 185, 216, 217, 222, 223,
 224, 225, 227, 229, 230, 231, 233, 443,
 445, 446, 450, 452, 453, 476, 482
 Loyd, B. H., 449
 Luecht, R. M., 35
 Lumsden, J., 12, 228
 Lunn, D., 227
- MacCallum, R. C., 45, 376
 Mächler, M., 531, 541
 Mack, J. S., 491
 Magis, D., 151, 213, 520
 Maier, K. S., 525
 Maier, M. J., 44
 Mair, P., 44, 74
 Makransky, G., 191
 Mantel, N., 483
 Marais, I., 190, 191
 Maranell, G. M., 3
 Marco, G. L., 449
 Masters, G. N., 20, 37, 60, 238, 240, 241, 244,
 267, 272, 277, 297, 299, 313, 314, 324, 329,
 351
 Maydeu-Olivares, A., 121, 421
 McCoach, D. B., 531
 McCullagh C. E., 588
 McDonald, R. P., 11, 38, 48, 52, 102, 227, 405,
 406, 408
 McKinley, R. L., 98, 129, 393, 400, 401, 405,
 406, 407, 440, 590
 McLaughlin, M. E., 209
 McMahon, J. M., 589
 Meijer, R. R., 209, 213
 Mellenbergh, G. J., 364, 482
 Meulders, M., 567, 588
 Miller, T. R., 490
 Mills, C. N., 98, 129, 590
 Mislevy, R. J., 45, 78, 88, 92, 93, 95, 96, 97,
 98, 102, 105, 107, 173, 183, 187, 222, 223,
 436, 440, 557, 588
 Miyazaki, Y., 525
 Mok, M., 38
 Mooney, J. A., 188
 Muckle, T. J., 530
 Muraki, E., 45, 301, 305, 313, 314, 351, 364
- Muthén, B. O., 143
 Muthén, L. K., 143
- Nagelkerke, N. J. D., 490, 524
 Nanda, H., 569
 Nandakumar, R., 39
 Nering, M. L., 189, 212
 Neve, H. J., 588
 Neyman, J., 87
 Niessen, A. S. M., 213
 Nieto, R., 530
 Novick, M. R., 6, 17, 22, 38, 169, 224, 230, 231
 Nungester, R. J., 547
 Nunnally, J. C., 3
- O'Connell, A. A., 589
 Orlando, M., 159, 189, 253
 Osborne, D., 237
 Oshima, T. C., 188, 222, 454
- Panter, A. T., 48
 Pastor, D. A., 525, 587
 Patterson, D. G., 344, 354
 Patz, R. J., 126
 Pearson, J. S., 237
 Pedhauzer, E. J., 520
 Peng, C. Y. J., 589
 Peres, D., 568
 Petersen, N. S., 445, 475
 Pimentel, J. L., 223
 Pinheiro, J., 531
 Plake, B., 223
 Pollack, J. M., 185
 Pouget, E. R., 589
- Rabe-Hesketh, S., 588
 Raïche, G., 213
 Rajaratnam, N., 569
 Raju, N. S., 150, 482
 Ramsay, J. O., 36
 Randall, J., 525
 Rasbash, J., 531
 Rasch, G., 16, 27, 37, 40, 78
 Rasinski, K. A., 185
 Raudenbush, S. W., 531, 588
 Reckase, M. D., 172, 392, 393, 398, 400, 401,
 405, 406, 407, 411, 439, 440, 442, 447
 Ree, M. J., 447, 449
 Reise, S. P., 98, 136, 293, 333
 Revuelta, J., 392
 Ricker, K. L., 446
 Rijmen, F., 229, 438, 568

- Roberts, J. K., 536
 Robitzsch, A., 44, 48, 515
 Rock, D. A., 185
 Rogers, H. J., 39, 589
 Rogers, J., 482, 488, 489, 508
 Rose, N., 223
 Rosenbaum, P. R., 189
 Roskam, E. E., 22
 Rost, J., 442
 Roth, R. L., 222
 Rubin, D. B., 92, 221, 444
 Rupp, A. A., 67, 68
- Saisangjan, U., 224
 Samejima, F., 31, 77, 185, 186, 294, 311, 312,
 324, 328, 329, 333, 334, 335, 346, 347,
 349, 350–351, 352, 354, 364, 367, 385,
 440
 Sarkar D., 531
 Sava-Bolesti, M., 97, 368, 369
 Schaeffer, N. C., 136
 Schafer, W. D., 97
 Scheers, N. J., 8
 Scheffé, H., 87
 Schepers, J., 568
 Schumacker, R. E., 58
 Schwarz, G., 152
 Schwarz, R. D., 392, 407
 Sclove, S. L., 159
 Scott, E. L., 87
 Scott, L. A., 185
 Searle, S. R., 588
 Secrest, D., 127
 Seliger, E., 436
 Seong, T.J., 93, 97, 141
 Shepard, L. A., 174, 479, 482, 485
 Shojima, K., 351
 Shultz, M., 292
 Sijtsma, K., 209, 364
 Silk, K. J., 558
 Sinharay, S., 446
 Sireci, S. G., 23, 188
 Skaggs, G., 229, 525
 Skrondal, A., 588
 Smith, E. V., 74
 Smith, R. M., 58, 61, 62
 Snell, E. J., 524
 Snijders, T. A. B., 213
 Spiegelhalter, D., 227
 Spiegelhalter, D. J., 159
 Spray, J. A., 392
 Steinberg, L., 188, 301, 333, 364, 367, 482,
 486
- Stenner, A. J., 74
 Stevens, S. S., 1, 4, 9
 Stevenson, J., 229
 Stocking, M. L., 93, 107, 223, 450, 452
 Stone, C. A., 141, 159, 176, 376, 385
 Stone, M. H., 16, 20, 39, 43, 74
 Stouffer, S. A., 8
 Stout, W., 38, 39
 Stroud, A. H., 126
 Subkoviak, M. J., 491
 Sugawara, H. M., 376
 Sugiura, N., 152
 Sulis, I., 587
 Swaminathan, H., 39, 87, 95, 114, 185, 229,
 482, 488, 489, 508
 Swanson, D. B., 547
 Swenson, W. M., 237
 Swygert, K. A., 48
 Sykes, R. C., 186
 Sympson, J. B., 392, 436
- Tanaka, J. S., 48, 52
 Tatsuoka, K. K., 357
 Tendeiro, J. N., 213
 Thayer, D. T., 444, 482, 483, 485, 486
 Theunissen, T. J. J. M., 35
 Thissen, D., 45, 159, 188, 189, 190, 191, 252,
 253
 Thissen, D. J., 88, 93, 141, 142, 187, 188, 223,
 225, 229, 248, 293, 301, 361, 364, 365,
 366, 367, 370, 376, 385, 388, 390, 482,
 486, 487, 488, 520
 Thomas, A., 227
 Thomson, William (Lord Kelvin), 1
 Thurstone, L. L., 3, 8, 10, 12, 48, 401
 Tibshirani, R. J., 421
 Toland, M. D., 587
 Torres Irribarra, D., 74
 Tryon, R. C., 569
 Tucker, L. R., 17
 Tuerlinckz, F., 151, 520
 Tutz, G., 297
- van Buuren, S., 223
 Van den Noortgate, W., 547, 567, 588
 van den Wollenberg, A., 98
 van der Linde, A., 159
 van der Linden, W. J., 35
 van der Wollenberg, A. L., 93
 VanLehn, K., 357
 van Nispen, R. M. A., 588
 van Rens, G. H. M. B., 588
 Verhelst, N. D., 22, 44, 98, 225, 307

- Vermunt, J. K., 579
Verstralen, H. H. F. M., 22, 44, 307
von Davier, A. A., 444, 446
von Davier, M., 223, 225
- Wainer, H., 127, 142, 183, 187, 188, 189, 223,
224, 225, 229, 479, 482, 486
Walker, C. M., 525
Walker, D. A., 39, 77
Walker, S. C., 531, 541
Walker-Barnick, L. A., 292
Wallace, S., 3
Waller, N. G., 136
Walsh, D., 8
Wang, M. D., 439
Wang, S., 525, 578
Wang, T., 349
Wang, W., 525
Wang, W. C., 189
Wang, X., 189
Wang, X. B., 223
Warm, T. A., 127
Way, W. D., 439
Weeks, J. P., 450
Weitzman, R. A., 127
Whitely (Embretson), S. E., 45, 392, 436
Wickham, H., 541, 582
Widaman, K. F., 45
Williams, D. M., 174
Williams, E. A., 162, 209
Willse, J., 46, 287
- Wilson, M., 5, 20, 189, 244, 299, 525, 527, 574,
587, 588
Wilson, M. R., 78
Wingersky, M. S., 44, 184
Wollack, J. A., 222, 369
Wright, B. D., 16, 20, 37, 39, 43, 44, 45, 56, 58,
59, 60, 62, 65, 74, 83, 84, 122, 127, 224,
228, 244, 272, 277, 278, 281, 283, 292, 310
Wu, M., 44, 527, 588
Wu, M. L., 78
Wu, P., 222, 223
- Xu, X., 223
- Yao, L., 392, 407
Yen, W. M., 12, 185, 186, 187, 188, 189, 190,
191, 225, 235, 314, 475, 590
Yu, J., 293, 333
Yu, K. F., 87
- Zeng, L., 41, 349, 444, 450
Zhang, B. J., 159, 176
Zhang, S., 3, 45
Zieky, M., 482
Zimowski, M., 45
Zoellner, J., 558
Zoplouoglu, C., 351
Zumbo, B. D., 67, 68, 479, 490, 523
Zwarts, M., 92
Zwinderman, A. H., 37, 93

Subject Index

Note. *t* and *f* indicate tables and figures.

- Akaike information criterion (AIC)
application of the 2PL model using MMLE and mirt, 152, 159
application of the RS model to an attitudes towards condoms scale, JMLE, mixRasch, 288
three-parameter logistic (3PL) model and, 179, 199–200, 199t
AMT-Robustified Jackknife MLE, 127
Analysis of variance (ANOVA), 21, 87–88, 176
Anchor test, 446
Appropriateness measurement, 209, 211–216, 214t, 215f, 216f
Assumptions, 6–7, 10–11, 21–23
- Bayes Mean Estimate. *See* Expected a posteriori (EAP) method
Bayes Modal Estimate. *See* Maximum a posteriori (MAP)
Bayesian information criterion (BIC)
application of the 2PL model using MMLE and mirt, 152, 159
application of the RS model to an attitudes towards condoms scale, JMLE, mixRasch, 288
three-parameter logistic (3PL) model and, 179, 199–200, 199t
Behavioral manifestations, 2, 9–10
Bifactor model, 438, 439f
BIGSTEPS program
application of the Rasch model to the mathematics data, JMLE, BIGSTEPS, 46–68, 49t–51t, 55t, 57t, 61t, 63t, 65t, 66t
application of the RS model to an attitudes towards condoms scale, JMLE, BIGSTEPS, 272–287, 274t, 275t–276t, 277f, 279t–280t, 282t, 284t–285t, 285f
- calibration and, 53–56, 55t, 57t
dimensionality assessment and, 47–53, 49t–51t
location estimates and, 83–84
model–data fit assessment, 56–64, 61t, 63t
person location estimation and, 64–68, 66t
BILOG-MG
application of the 2PL model to the mathematics data, MMLE, BILOG-MG, 143–146, 143t, 145t–146t, 147t–148t, 148f
application of the 3PL model to the mathematics data, MMLE, BILOG-MG, 192–195, 192f, 193t–194t, 195f
application of the 3PL model using MMLE and BILOG-MG, 192–195, 192t, 193t–194t, 195f, 196f
application of the Rasch model to the mathematics data, MMLE, BILOG-MG, 98–109, 99t, 101t, 103t–104t, 106f, 107f, 110t
metric transformation and, 112–113
overview, 127, 128–129
two-parameter logistic (2PL) model and, 173–174
Biserial correlation coefficient, 127–128
Bock–Samejima model for multiple-choice items (BS), 367, 388–389. *See also* Multiple-choice (MC) model
Boundary characteristics curves. *See* Category boundary location
- Calibration
equating and, 447
metric transformation and, 112–113
missing data and, 223–224
multidimensional calibration, 442
overview, 45–46, 78

- Calibration (*cont.*)
 Rasch model and, 77
 sample size and, 140–142, 187–188, 292–294, 293f, 333–334, 368–370
 three-parameter logistic (3PL) model and, 185, 229–230
- Category boundary location, 328–330, 330f
- Category information functions, 294–296
- Category probability curves. *See* Option response function (ORF)
- Category response functions. *See* Option response function (ORF)
- Category scores, 238, 296–297
- Centering approaches, 44, 77–78
- Chi-square statistic, 176–177, 483–486, 590–591
- Classical test theory (CTT), 4, 5–8
- Close response situation, 354–355
- Cochran–Mantel–Haenszel (CMH) statistic.
See also Mantel–Haenszel (MH) statistic
- DIF analysis of vocabulary test, *mantelhaen.test* and *difR*, 494–501, 495t–496t, 497t–500t, 501f, 502f
 - DIF analysis of vocabulary test, SAS CMH and, 491–493, 492t, 493f
- Common-item groups design, 445
- Common-item nonequivalent groups design, 445–446
- Compensatory models, 392, 437–438, 439f.
See also Multidimensional item response theory (MIRT) model
- Conceptual parameter estimation, 139–140, 140f, 243–244. *See also* Parameter estimates
- Concurrent calibration, 447, 448f, 465, 468–471, 469t–471t, 472t–474t. *See also* Calibration
- Conditional independence assessment, 195–198, 196t, 197f
- Conditional independence assumption overview, 21–22
 principles of, 38–39
- three-parameter logistic (3PL) model and, 188–192, 195–198, 196t, 197f, 230
- Conditional maximum likelihood estimation (CMLE), 43–44. *See also* Likelihood function; Parameter estimates
- Conditional reliability, 172–173
- Conditional standard error of measurement, 40–41. *See also* Standard error of estimate (SEE)
- Conformal patterns, 39
- Consistency of the measures. *See* Reliability
- Construct-related validity evidence, 75–76
- Content validity, 75–76
- Content-oriented factors, 47–48
- Continuous rating scale, 354
- Continuous response model, 333, 343–351, 348f, 350f
- Convergence criterion, 43
- Counterbalanced random-groups data collection method, 445
- Cramér's V, 262–264, 264t
- Criterion-related validity evidence, 75–76
- Cumulative probability curves. *See* Category boundary location
- Curvilinearity, 47–48
- Data calibration. *See* Calibration
- Data collection, 445–446, 474
- Dependence, 188–192
- Design matrix, 588–589
- Deviance information criterion (DIC), 159
- Dichotomized polytomous responses, 356–357
- Dichotomous item response, 526–530, 528f
- Dichotomous model approach, 189, 225–226
- Dichotomous Rasch model, 68–69
- Differential item functioning (DIF) analysis
- DIF analysis of vocabulary test, *glm* and *difR*, 508, 511–515, 512t–514t, 516t–518t
 - DIF analysis of vocabulary test, *mantelhaen.test* and *difR*, 494–501, 495t–496t, 497t–500t, 501f, 502f
 - DIF analysis of vocabulary test, SAS CMH, 491–493, 492t, 493f
 - DIF analysis of vocabulary test, SAS proc logistic, 501–508, 502t, 504t–505t, 507t, 509t–510t, 511f
- item bias and, 479–482, 480t, 481f
- logistic regression, 487–490
- Mantel–Haenszel (MH) statistic and, 483–486
- overview, 76, 475, 478–479, 518–519
- person-level predictors for items–DIF analysis, 1me4, 551–556, 552t–555t
- person-level predictors for items–DIF analysis, proc glimmix, 547–550, 547t
- TSW likelihood ratio test, 486–487
- difR* package
- DIF analysis of vocabulary test, *glm* and *difR*, 508, 511–515, 512t–514t, 516t–518t

- DIF analysis of vocabulary test, *mantelhaen.test* and *difR*, 494–501, 497t–500t, 501f, 502f
- Dimensionality assessment, 47–53, 49t–51t, 83
- EM algorithm, 92–93, 102
- Empirical reliability, 144, 146
- EQUATE program, 454–461, 456t–459t, 460t, 461f, 462t–463t, 464f, 465f
- Equating
- application of the total characteristic function equating method, EQUATE, 454–461, 456t–459t, 460t, 461f, 462t–463t, 464f, 465f
 - application of the total characteristic function equating method, SNSequate and, 463–465, 466t–467t
 - data collection phase, 445–446
 - fixed-item and concurrent calibration equating, 465, 468–471, 469t–471t, 472t–474t
 - horizontal equating, 475
 - overview, 443–444, 471, 474–475, 476–477
 - transformation phase, 446–454, 448f, 451f, 452f, 454f
- Equating coefficients, 447–448. *See also* Metric transformation
- Equipercentile equating, 444. *See also* Equating
- Essential dimensionality, 38–39
- Essential independence (EI), 38–39
- E-step, 92, 93. *See also* EM algorithm
- Expected a posteriori (EAP) method
- application of the 2PL model using MMLE and BILOG-MG, 144, 146
 - application of the GR model to an attitudes toward condoms scale, MMLE, *mirt*, 340, 343
 - missing data and, 223
 - overview, 86, 93–98, 94f, 97t, 125, 173
 - three-parameter logistic (3PL) model and, 185
- Expected trait score, 6, 114
- External common items variant, 446
- Factor analytic model, 392–393, 588
- Fit analysis. *See also* Model–data fit
- three-parameter logistic (3PL) model and, 179, 198–200, 199t
 - two-parameter logistic (2PL) model and, 146, 148–151, 148f, 151f
- Fixed-item parameter approach, 447, 465, 468–471, 469t–471t, 472t–474t
- flexMIRT
- application of the GPC model to a reasoning ability instrument, MMLE, flexMIRT, 318–320, 320t, 321t, 322t, 323t–324t
 - application of the GR model to an attitudes toward condoms scale, MMLE, flexMIRT, 337–338, 338t, 339t
 - application of the PC Model to a reasoning ability instrument, MMLE, flexMIRT, 244–256, 247t, 249t–252t, 254f, 255f, 256f
 - calibration of interpersonal engagement instrument, M2PL model, flexMIRT and, 429–431, 430f, 431f, 432t–436t
 - calibration of interpersonal engagement instrument, M2PL model, *sirt.noharm* and, 411–421, 413t–418t, 419t, 420t, 421f
 - overview, 299, 307–310
- Four-parameter logistic model (4PL), 227
- Full-information factor analysis (FIFA), 440
- Functional form assumption, 22–23
- Fundamental measurement, 167–168
- Gaussian error function, 127
- General linear mixed models. *See* Multilevel IRT models
- Generalized linear mixed (effects) models. *See* Multilevel IRT models
- Generalized linear mixed models. *See* Multilevel IRT models
- Generalized linear mixed models (GLMM), 588
- Generalized linear model (*glm*) function
- DIF analysis of vocabulary test, *glm* and *difR*, 508, 511–515, 512t–514t, 516t–518t
- overview, 588
- Rasch model estimation, *lme4*, 541–545, 542t–543t
- Generalized partial credit (GPC) model. *See also* Partial credit (PC) model; Polytomous model approach
- application of the GPC model to a reasoning ability instrument, MMLE, flexMIRT, 318–320, 320t, 321t, 322t, 323t–324t
 - application of the GPC model to a reasoning ability instrument, MMLE, *mirt*, 321–322, 324, 325t–327t, 328f
 - metric transformation and, 336–337
 - nominal response (NR) model and, 364–365

- Generalized partial credit (GPC) model (*cont.*)
 overview, 313–318, 315f, 316f, 317f, 318f,
 319f, 351–352
- Goodness-of-fit index (GFI), 52–53
- Graded response (GR) model. *See also* Polytomous model approach
 application of the GR model to an attitudes toward condoms scale, MMLE, flexMIRT, 337–338, 338t, 339t
- application of the GR model to an attitudes toward condoms scale, MMLE, mirt, 340–343, 341t–343t, 344f, 345f, 346f
- calibration sample size and, 333–334
- conceptual development of, 324, 328–333, 330f, 331f, 332f
- dichotomous data and, 352–353
- information for graded data, 334–336, 336f
- metric transformation and, 336–337
 overview, 313, 351–352
- Graphic response scale, 354
- Guttman scale, 39, 84
- Guttman Scalogram technique, 38
- Half-item rule, 84
- Hermite–Gauss quadrature method, 89–90, 90f, 127
- Heterogeneous GR model, 333. *See also* Graded response (GR) model
- Hierarchical linear models (HLMs). *See also* Multilevel IRT models
- Homogeneity of variance assumption in ANOVA, 21
- Horizontal equating, 475. *See also* Equating
- Ideal response patterns, 39
- Imputation, 221–222
- Indeterminacy, 408–410, 409f, 446–447
- Indeterminacy of scale, 44–45, 77–78
- INFIT statistic. *See also* Model–data fit
 overview, 57–59, 78, 84–85
 polytomous model approach, 293f
 Rasch model and, 77
- Inflexion line, 395
- Information available for estimation
 graded response (GR) model and, 334–336, 336f
 nominal response (NR) model and, 365–366
 for the PC and RS models, 294–296
 three-parameter logistic (3PL) model and, 216–220, 218f, 219f, 230–233
 two-parameter logistic (2PL) model and, 162–165, 164f, 166f
- Information criterion, 288. *See also* Akaike information criterion (AIC); Bayesian information criterion (BIC)
- Initial metric, 447–448
- Internal common items variant, 445–446
- Invariance, 3–4, 67–68
- Item bias, 479–482, 480t, 481f
- Item centering, 44, 77–78
- Item characteristic curve (ICC), 17
- Item characteristic function, 17
- Item cluster, 188–189
- Item curve, 17
- Item difficulties, 16
- Item discrimination parameter
 metric transformation and, 112–113
 overview, 19, 228
 partial credit model and, 304
 three-parameter logistic (3PL) model and, 187, 219–220
- two-parameter logistic (2PL) model and, 149
- Item information area index, 162–165, 164f, 166f, 216–220, 218f, 219f, 230–233
- Item information function
 application of the PC Model to a reasoning ability instrument, MMLE, flexMIRT, 254–255, 254f
 graded response (GR) model and, 334–336, 336f
 PC and RS models and, 294–296
- Item interdependency, 22
- Item location estimation
 estimation capacity and, 32–35, 33f, 34f
 multilevel IRT models and, 526–527
 overview, 23–28, 24f, 26f, 27f, 36, 77–78, 84, 87
 rating scale (RS) model and, 267–268, 268f
 standard error of estimate (SEE) and, 31
- Item operating characteristic, 17
- Item parameter characterization, 267–268, 267f
- Item parameter drift, 520
- Item parameter estimation, 304, 477
- Item removal, 76–77
- Item response function (IRF)
 estimation capacity and, 32–35, 33f, 34f, 38
 joint maximum likelihood estimation (JMEL) and, 84
 logistic regression and, 487–490
 one-parameter model and, 19–20
 overview, 17, 18f, 41, 41f
 three-parameter logistic (3PL) model and, 183–186, 184f
 two-parameter logistic (2PL) model and, 165, 167, 168–169

- Item response surface (IRS), 394–395, 394f, 395f, 396f, 406–407, 407f
- Item response vectors, 84
- Item vectors, 401–404, 402f, 439–440
- Item-level fit, 60–64, 61t, 63t, 252–253
- Item-level model, 527
- Item-level predictors for items
- item-level predictors for items–nutrition literacy, lme4, 571, 574, 575t–577t
 - item-level predictors for items–nutrition literacy, proc glmmix, 569–571, 570f, 571t, 572t–573t
 - overview, 567–569, 587
- Item-Person Map. *See* Variable Map
- Joint Distribution Map. *See* Variable Map
- Joint maximum likelihood estimation (JMLE). *See also* Likelihood function; Parameter estimates
- application of the Rasch model to the mathematics data, JMLE, BIGSTEPS, 46–68, 49t–51t, 55t, 57t, 61t, 63t, 65t, 66t
 - application of the Rasch model to the mathematics data, JMLE, mixRasch, 68–74, 70t–72t, 73f
 - application of the RS model to an attitudes towards condoms scale, JMLE, BIGSTEPS and, 272–287, 274t, 275t–276t, 277f, 279t–280t, 282t, 284t–285t, 285f
 - application of the RS model to an attitudes towards condoms scale, JMLE, mixRasch, 287–292, 289t–291t, 292f
 - calibration and, 45–46, 187–188
 - estimation of the M2PL model's item parameters and, 405–406
 - indeterminacy of scale and, 45
 - overview, 42–44, 77–78, 86–88, 125, 171
 - Rasch model and, 76–77
 - rating scale (RS) model and, 272
 - three-parameter logistic (3PL) model and, 229
 - validity evidence and, 75–76
- j* 's multidimensional item location, 397–401
- Latent class analysis (LCA), 4, 8–9
- Latent class model, 8, 588
- Latent structure analysis (LSA), 11
- Latent variables
- item interdependency and, 22
 - latent variable continuum, 2, 12–13, 13f
 - overview, 2, 5, 10
 - partial credit model and, 238–239
- Likelihood function, 25–28, 26f, 27f, 28f, 42–43, 85, 185–186, 186f. *See also* Conditional maximum likelihood estimation (CMLE); Joint maximum likelihood estimation (JMLE); Maximum likelihood estimation (MLE)
- Likelihood ratio statistic, 85, 179, 488–490
- Likert response scale data, 267. *See also* Rating scale (RS) model
- Linear equating, 444. *See also* Equating
- Linear Logistic Test Model (LLTM)
- item-level predictors for items–nutrition literacy, proc glmmix, 569–571, 570f, 571t, 572t–573t
 - multilevel IRT models and, 595–596
 - overview, 5
- Linear transformation, 111, 162–165, 164f, 166f
- Linking, 443, 471, 474–475
- lme4 package
- item-level predictors for items–nutrition literacy, lme4, 571, 574, 575t–577t
 - person-level predictors for items–DIF analysis, lme4, 551–556, 552t–555t
 - person-level predictors for respondents–nutrition literacy, lme4, 562–567, 563t–566t
 - Rasch model estimation, lme4, 541–545, 542t–543t
 - three-level model analysis–nutrition literacy, lme4, 582–587, 583t–586t
- Local independence. *See* Conditional independence assumption
- Location estimation
- joint maximum likelihood estimation (JMLE) and, 83–84
 - nominal response (NR) model and, 364–365
 - overview, 23–28, 24f, 26f, 27f
- Log likelihood functions, 25–28, 26f, 27f, 28f, 139–140, 140f
- Log odds, 358–359
- Logistic function, 392–393
- Logistic regression
- DIF analysis of vocabulary test, glm and difR, 508, 511–515, 512t–514t, 516t–518t
 - DIF analysis of vocabulary test, SAS CMH, 491–493, 492t, 493f
 - DIF analysis of vocabulary test, SAS proc logistic, 501–508, 502t, 504t–505t, 507t, 509t–510t, 511f
 - overview, 487–490, 519
- Logit regression line, 18, 361–362, 362f

- Logit regression plane, 396, 397*f*
 Logit transformation, 357–359, 358*f*
- Manifest variables, 2, 9–10
 Mantel–Haenszel (MH) statistic
 DIF analysis of vocabulary test, *mantelhaen.test* and *difR*, 494–501, 495*t*–496*t*, 497*t*–500*t*, 501*f*, 502*f*
 DIF analysis of vocabulary test, SAS CMH and, 491–493, 492*t*, 493*f*
 DIF analysis of vocabulary test, SAS *proc logistic*, 501–508, 502*t*, 504*t*–505*t*, 507*t*, 509*t*–510*t*, 511*f*
 overview, 483–486, 519
 in SPSS, 520–524
`mantelhaen.test` function, 494–501, 495*t*–496*t*, 497*t*–500*t*, 501*f*, 502*f*
 Many-Facet Rasch Model (MFRM), 306–307
 Marginal maximum likelihood estimation (MMLE). *See also* Parameter estimates
 application of the 2PL model to the mathematics data, MMLE, BILOG-MG, 143–146, 143*t*, 145*t*–146*t*, 147*t*–148*t*, 148*f*
 application of the 2PL model to the mathematics data, MMLE, *mirt*, 152–162, 153*t*–158*t*, 160*f*, 161*f*
 application of the 2PL model using MMLE and BILOG-MG, 143–146, 143*t*, 145*t*–146*t*, 147*t*–148*t*
 application of the 2PL model using MMLE and *mirt*, 152–162, 153*t*–158*t*, 160*f*, 161*f*
 application of the 3PL model to the mathematics data, MMLE, BILOG-MG, 192–195, 192*f*, 193*t*–194*t*, 195*f*
 application of the 3PL model to the mathematics data, MMLE, *mirt*, 200–209, 201*t*–202*t*, 203*f*, 204*t*–208*t*, 209*f*, 210*f*, 211*f*, 212*f*
 application of the 3PL model using MMLE and BILOG-MG, 192–195, 192*t*, 193*t*–194*t*, 195*f*, 196*f*
 application of the 3PL model using MMLE and *mirt*, 200–209, 201*t*–202*t*, 203*f*, 204*t*–208*t*, 209*f*, 210*f*, 211*f*, 212*f*
 application of the GPC model to a reasoning ability instrument, MMLE, flexMIRT, 318–320, 320*t*, 321*t*, 322*t*, 323*t*–324*t*
 application of the GPC model to a reasoning ability instrument, MMLE, *mirt*, 321–322, 324, 325*t*–327*t*, 328*f*
 application of the GR model to an attitudes toward condoms scale, MMLE, flexMIRT, 337–338, 338*t*, 339*t*
- application of the GR model to an attitudes toward condoms scale, MMLE, *mirt*, 340–343, 341*t*–343*t*, 344*f*, 345*f*, 346*f*
 application of the NR model to a general science test, MMLE, *mirt*, 370–382, 371*t*–374*t*, 375*f*, 377*f*, 378*f*, 379*f*, 380*t*–381*t*, 382*f*, 383*f*
 application of the PC Model to a reasoning ability instrument, MMLE, flexMIRT, 244–256, 247*t*, 249*t*–252*t*, 254*f*, 255*f*, 256*f*
 application of the PC Model to a reasoning ability instrument, MMLE, *mirt*, 256–266, 257*t*–262*t*, 263*f*, 264*t*, 265*f*
 application of the Rasch model to the mathematics data, MMLE, BILOG-MG, 98–109, 99*t*, 101*t*, 103*t*–104*t*, 106*f*, 107*f*, 110*t*
 application of the Rasch model to the mathematics data, MMLE, *mirt*, 115–124, 116*t*–120*t*, 122*f*, 125*f*
 calibration sample size and, 141–142, 187–188
 expected a posteriori method and, 93–98, 94*f*, 97*t*
 metric transformation and, 111–114, 115*f*
 overview, 85–93, 90*f*, 125–126, 171, 172
 rating scale (RS) model and, 272
 total characteristic function and, 111–114, 115*f*
 Marginal reliability, 172–173
 Markov chain Monte Carlo (MCMC) simulation methods, 126. *See also* Monte Carlo simulations
 Masters (Rasch) PC model, 299–301, 302*t*–303*t*. *See also* Partial credit (PC) model
 Maximum a posteriori (MAP), 95–96, 125, 320, 323*t*–324*t*
 Maximum likelihood estimation (MLE). *See also* Joint maximum likelihood estimation (JMLE); Marginal maximum likelihood estimation (MMLE)
 estimation capacity and, 32–35, 33*f*, 34*f*, 40
 expected a posteriori method and, 95–96
 missing data and, 222, 223
 overview, 27–28, 88, 125, 172
 pragmatic characteristics of, 28–29, 28*f*
 three-parameter logistic (3PL) model and, 185, 233–234
 Mean equating, 444. *See also* Equating
 Mean-mean approach, 449
 Mean-sigma method, 449
 Measurement, 1–4, 9

- Metric indeterminacy, 408, 446–447. *See also* Indeterminacy
- Metric transformation. *See also* Transformations
- GPC and GR models and, 336–337
 - metric transformation coefficients, 111–114, 115f, 447–448
 - nominal response (NR) model and, 366
 - overview, 449–450, 474–475
 - PC and RS models and, 296
 - three-parameter logistic (3PL) model and, 220, 221f
 - two-parameter logistic (2PL) model and, 142–143
- mirt* package
- application of the 2PL model to the mathematics data, MMLE, *mirt*, 152–162, 153t–158t, 160f, 161f
 - application of the 2PL model using MMLE and *mirt*, 152–162, 153t–158t, 160f, 161f
 - application of the 3PL model to the mathematics data, MMLE, *mirt*, 200–209, 201t–202t, 203f, 204t–208t, 209f, 210f, 211f, 212f
 - application of the 3PL model using MMLE and *mirt*, 200–209, 201t–202t, 203f, 204t–208t, 209f, 210f, 211f, 212f
 - application of the GPC model to a reasoning ability instrument, MMLE, *mirt*, 321–322, 324, 325t–327t, 328f
 - application of the GR model to an attitudes toward condoms scale, MMLE, *mirt*, 340–343, 341t–343t, 344f, 345f, 346f
 - application of the NR model to a general science test, MMLE, *mirt*, 370–382, 371t–374t, 375f, 377f, 378f, 379f, 380t–381t, 382f, 383f
 - application of the PC Model to a reasoning ability instrument, MMLE, flexMIRT, 244–256, 247t, 249t–252t, 254f, 255f, 256f
 - application of the PC Model to a reasoning ability instrument, MMLE, *mirt*, 256–266, 257t–262t, 263f, 264t, 265f
 - application of the Rasch model to the mathematics data, MMLE, *mirt*, 115–124, 116t–120t, 122f, 125f
 - calibration of interpersonal engagement instrument, M2PL model, flexMirt, 429–431, 430f, 431f, 432t–436t
 - calibration of interpersonal engagement instrument, M2PL model, *mirt*, 422–429, 423t–428t
 - nominal response (NR) model and, 389–390
 - overview, 133–134, 134f
 - Missing at random (MAR) data, 221
 - Missing by design, 222
 - Missing completely at random (MCAR) data, 221
 - Missing data, 220–224
 - Missing not at random (MNAR) data, 221
 - mixRasch* package
 - application of the Rasch model to the mathematics data, JMLE, *mixRasch*, 68–74, 70t–72t, 73f
 - application of the RS model to an attitudes towards condoms scale, JMLE, *mixRasch*, 287–292, 289t–291t - Mixture models, 4, 225
 - Model identification problem. *See* Indeterminacy of scale
 - Model–data fit. *See also* Fit analysis
 - issues to consider in selecting among the 1PL, 2PL, and 3PL models, 224–226
 - MMLE and, 125
 - overview, 77–78, 167–168, 478
 - person parameter estimation and, 87
 - Rasch model and, 76–77
 - three-parameter logistic (3PL) model and, 179, 198–200, 199t
 - two-parameter logistic (2PL) model and, 146, 148–151, 148f, 151f
 - using invariance for, 67–68 - Monte Carlo simulations, 141, 187. *See also* Markov chain Monte Carlo (MCMC) simulation methods
 - M-step, 92. *See also* EM algorithm
 - MSTEPS, 292–293
 - Multidimensional generalizes partial credit (MGPC), 392
 - Multidimensional graded response (MGR), 392
 - Multidimensional item difficulty, 398
 - Multidimensional item information surface, 406–407, 407f
 - Multidimensional item location, 397–401
 - Multidimensional item response theory (MIRT) model. *See also* Multidimensional three-parameter logistic (M3PL) model; Multidimensional two-parameter logistic (M2PL) model
 - assumptions of, 404–405
 - conceptual development of, 391–396, 394f, 395f, 396f, 397f
 - indeterminacy in, 408–410, 409f
 - item vectors and vector graphs and, 401–404, 402f

- Multidimensional item response theory (MIRT) model (*cont.*)
- multidimensional item location and discrimination, 397–401
 - overview, 391, 431, 436–437
 - person location estimation and, 421–422
- Multidimensional nominal response (MNR), 392
- Multidimensional three-parameter logistic (M3PL) model, 404, 436–437. *See also* Multidimensional item response theory (MIRT) model; Three-parameter logistic (3PL) model
- Multidimensional two-parameter logistic (M2PL) model. *See also* Multidimensional item response theory (MIRT) model;
- Two-parameter logistic (2PL) model
 - calibration of interpersonal engagement instrument, M2PL model, flexMirt, 429–431, 430f, 431f, 432t–436t
 - calibration of interpersonal engagement instrument, M2PL model, mirt, 422–429, 423t–428t
 - calibration of interpersonal engagement instrument, M2PL model, sirt.noharm, 411–421, 413t–418t, 419t, 420t, 421f
 - estimation of, 405–406
 - information for, 406–407, 407f
 - overview, 393–394, 436–437
- Multidimensional two-parameter normal ogive model, 405–406
- Multilevel generalized linear mixed models. *See* Multilevel IRT models
- Multilevel IRT models
- estimating the Rasch model from a multilevel perspective, proc glimmix, 530–541, 532f, 533t, 534t–535t, 536t, 538t–540t
 - item-level predictors for items, 567–569
 - item-level predictors for items–nutrition literacy, lme4, 571, 574, 575t–577t
 - item-level predictors for items–nutrition literacy, proc glimmix, 569–571, 570f, 571t, 572t–573t
 - overview, 525, 587, 588
 - person-level predictors for items, 545–547
 - person-level predictors for items–DIF analysis, lme4, 551–556, 552t–555t
 - person-level predictors for items–DIF analysis, proc glimmix, 547–550, 547t, 548t–551t
 - person-level predictors for respondents, 556–557
- person-level predictors for respondents–nutrition literacy, lme4, 562–567, 563t–566t
- person-level predictors for respondents–nutrition literacy, proc glimmix, 558–562, 559t, 560t–561t, 562f
- Rasch model estimation, lme4, 541–545, 542t–543t
- three levels, 574, 577–579, 587
- three-level model analysis–nutrition literacy, lme4, 582–587, 583t–586t
- three-level model analysis–nutrition literacy, proc glimmix, 579–582, 579t, 580t–581t
- two levels, 525–530, 528f, 587
- Multimodal likelihood function, 186. *See also* Likelihood function
- Multinomial logit model, 384
- Multiple imputation (MI) methods, 222
- Multiple-choice (MC) model, 366–368
- Multiple-response model. *See* Multiple-choice (MC) model
- Newton–Gauss cycles, 102
- Newton–Raphson steps, 92–93
- NOHARM, 48, 49t–51t, 52, 79–84, 79t, 80f, 81t, 82t, 411–421, 413t–418t, 419t, 420t, 421f
- Nominal categories model. *See* Nominal response (NR) model
- Nominal polytomous data, 356–357
- Nominal response (NR) model
- application of the NR model to a general science test, MMLE, mirt, 370–382, 371t–374t, 375f, 377f, 378f, 379f, 380t–381t, 382f, 383f
 - calibration sample size and, 368–370
 - conceptual development of, 357–365, 358f, 362f, 363f
 - information for, 365–366
 - metric transformation and, 366
 - overview, 360, 383–384, 389–390
- Noncompensatory models, 392, 437–438, 440–442. *See also* Multidimensional item response theory (MIRT) model
- Nonignorable missing values, 221–222
- Nonoperational items, 476
- Nonresponses. *See* Missing data
- Nonuniform DIF, 480–481, 481f, 519. *See also* Differential item functioning (DIF) analysis
- Not-reached items, 222

- Omitted responses, 222–223
- One-parameter (1PL) logistic model
issues to consider in selecting, 224–226
metric transformation and, 112–113
overview, 12, 17–20, 18f, 19f, 35–36, 87
Rasch model and, 20–21
- Open response situation, 354–355
- Operating characteristic curves. *See* Option response function (ORF)
- Operational items, 2, 476
- Option characteristics curves. *See* Option response function (ORF)
- Option information functions, 294–296
- Option response function (ORF). *See also* Trace line
application of the NR model to a general science test, MMLE, mirt, 375–379, 375f, 377f, 378f, 382f
- application of the PC Model to a reasoning ability instrument, MMLE, flexMIRT, 254–255, 254f, 255f
- generalized partial credit (GPC) model and, 314–318, 315f, 316f, 317f, 318f, 319f
- graded response (GR) model and, 330–333, 331f, 332f
- nominal response (NR) model and, 362–363, 363f
- overview, 241–242, 244f, 296–297
- rating scale (RS) model and, 269–271, 270f, 271f
- Ordinal logistic regression models, 352
- Ordinal polytomous data, 226–227. *See also* Polytomous model approach
- OUTFIT statistic. *See also* Model–data fit
overview, 58–59, 78, 84–85
polytomous model approach, 293f
- Rasch model and, 77
- Parameter estimates. *See also* Conceptual parameter estimation; Conditional maximum likelihood estimation (CMLE); Joint maximum likelihood estimation (JMLE); Marginal maximum likelihood estimation (MMLE); Unconditional maximum likelihood estimation (UCON)
calibration and, 45–46
indeterminacy of, 44–45
overview, 77–78
three-parameter logistic (3PL) model and, 229
two-parameter logistic (2PL) model and, 139–140, 140f, 149
- Partial credit (PC) model. *See also* Generalized partial credit (GPC) model; Polytomous model approach
application of the PC Model to a reasoning ability instrument, MMLE, flexMIRT, 244–256, 247t, 249t–252t, 254f, 255f, 256f
- application of the PC Model to a reasoning ability instrument, MMLE, mirt, 256–266, 257t–262t, 263f, 264t, 265f
- calibration sample size and, 292–294, 293f
- conceptual development of, 238–243, 241f, 243f
- conceptual parameter estimation of, 243–244
- information for, 294–296
- metric transformation and, 296
- nominal response (NR) model and, 364–365
- overview, 226–227, 296–297, 351
- three-mode data and, 306–307
- Pearson correlation coefficient, 67–68, 189–190
- Pearson's chi-square, 590–591
- Person centering, 44, 77–78. *See also* Centering approaches
- Person fit measures, 209, 211–216, 214t, 215f, 216f
- Person location estimation
appropriateness measurement and, 209, 211–216, 214t, 215f, 216f
estimation capacity and, 32–35, 33f, 34f
expected a posteriori method and, 86, 93–98, 94f, 97t
- multidimensional item response theory (MIRT) model and, 421–422
- multilevel IRT models and, 526–527
- overview, 23–28, 24f, 26f, 27f, 36, 77–78, 84, 170–171
- standard error of estimate (SEE) and, 31
- three-parameter logistic (3PL) model and, 233
- Person-Item Map. *See* Variable Map
- Person-level predictors for items
overview, 545–547, 587
- person-level predictors for items–DIF analysis, lme4, 551–556, 552t–555t
- person-level predictors for items–DIF analysis, proc glimmix, 547–550, 547t, 548t–551t
- Person-level predictors for respondents
overview, 556–557, 587

- Person-level predictors for respondents (*cont.*)
 person-level predictors for respondents—
 nutrition literacy, *lme4*, 562–567,
 563t–566t
 person-level predictors for respondents—
 nutrition literacy, *proc glimmix*,
 558–562, 559t, 560t–561t, 562f
- Point-biserial, 127–128
- Polytomous data, 237
- Polytomous model approach, 189, 226–227,
 237, 293f. *See also* Generalized partial
 credit (GPC) model; Graded response
 (GR) model; Partial credit (PC) model;
 Rating scale (RS) model
- Posterior distribution, 94–95, 94f
- Prior distribution
 calibration sample size and, 141–142
 expected a posteriori method and, 93–98,
 94f, 97t
 overview, 94–95, 94f
 three-parameter logistic (3PL) model and,
 185
- proc glimmix* (SAS)
 estimating the Rasch model from a mul-
 tilevel perspective, *proc glimmix*,
 530–541, 532f, 533t, 534t–535t, 536t,
 538t–540t
- item-level predictors for items—nutrition lit-
 eracy, *proc glimmix*, 569–571, 570f,
 571t, 572t–573t
- multilevel IRT models and, 589–590
- person-level predictors for items—DIF anal-
 ysis, *proc glimmix*, 547–550, 547t,
 548t–551t
- person-level predictors for respondents—
 nutrition literacy, *proc glimmix*,
 558–562, 559t, 560t–561t, 562f
- three-level model analysis—nutrition lit-
 eracy, *proc glimmix*, 579–582, 579t,
 580t–581t
- proc logistic* (SAS), 501–508, 502t,
 504t–505t, 507t, 509t–510t, 511f
- Product-moment matrix, 80, 83
- Pseudo-chance parameter, 181
- Pseudo-random guessing response function,
 180, 182–183, 226
- Q₃**
 conditional independence assessment and,
 195–198, 196t, 197f, 230
 overview, 301, 304
 three-parameter logistic (3PL) model and,
 188–192, 226
- Quadrature concept, 89–90, 90f, 93, 102, 105,
 126–127
- Random coefficient models. *See* Multilevel IRT
 models
- Randomly equivalent groups design, 445
- Rasch model
 application of the Rasch model to the math-
 ematics data, *JMLE*, *BIGSTEPS*, 46–68,
 49t–51t, 55t, 57t, 61t, 63t, 65t, 66t
 application of the Rasch model to the math-
 ematics data, *JMLE*, *mixRasch*, 68–74,
 70t–72t, 73f
 application of the Rasch model to the
 mathematics data, *MMLE*, *BILOG-MG*,
 98–109, 99t, 101t, 103t–104t, 106f, 107f,
 110t
 application of the Rasch model to the math-
 ematics data, *MMLE*, *mirt*, 115–124,
 116t–120t, 122f, 125f
 estimating the Rasch model from a mul-
 tilevel perspective, *proc glimmix*,
 530–541, 532f, 533t, 534t–535t, 536t,
 538t–540t
- indeterminacy of parameter estimates and,
 44
- issues to consider in selecting among the
 1PL, 2PL, and 3PL models, 224–226
- one-parameter model and, 20–21
- overview, 12–17, 13f, 14f, 15t, 35–36, 37,
 76–77
- Rasch model estimation, *lme4*, 541–545,
 542t–543t
- Rating scale (RS) model. *See also* Polytomous
 model approach
 application of the RS model to an attitudes
 towards condoms scale, *JMLE*, *BIG-
 STEPS*, 272–287, 274t, 275t–276t, 277f,
 279t–280t, 282t, 284t–285t, 285f
- application of the RS model to an atti-
 tudes towards condoms scale, *JMLE*,
 mixRasch, 287–292, 289t–291t, 292f
- calibration sample size and, 292–294, 293f
- conceptual parameter estimation of, 272
- information for, 294–296
- metric transformation and, 296
- overview, 226–227, 267–272, 267f, 268f,
 270f, 271f, 296–297
- three-mode data and, 306–307
- Regression models, 526–527. *See also* Logistic
 regression
- Relative efficiency (RE), 163–165, 164f, 166f
- Relative position, 443–444

- Reliability, 3. *See also* Consistency of the measures
- Resampling, 176–177
- Root mean square error of approximation (RMSEA), 52–53, 248
- Root mean square residual (RMSR), 52–53
- Root mean squared difference (RMSD), 149–151, 167, 174–175, 187
- Rotational indeterminacy, 408–410, 409f, 440. *See also* Indeterminacy
- Sample size
- calibration and, 45–46, 140–142, 187–188, 292–294, 293f, 333–334, 368–370
 - parameter estimation and, 140–142
- SAS program. *See proc glimmix (SAS); proc logistic (SAS)*
- Scale, indeterminacy of, 44–45, 77–78
- Scale development, 76–77
- Scale shrinkage, 475
- Scaling, 3, 475
- Scalogram analysis, 38
- Simultaneous calibration. *See Concurrent calibration*
- Single group with counterbalancing data collection approach, 445
- sirt.noharm, 411–421, 413t–418t, 419t, 420t, 421f
- Slider scale, 354
- SNSequate package, 463–465, 466t–467t
- Speededness, 222
- Spiraling, 445
- SPSS program, 520–524
- Standard error of estimate (SEE), 29–32, 30f, 32f, 40
- Standard error of measurement, 7
- Standard item approach, 78
- Standardized root mean square residual (SRMR), 52–53
- Strong principle of conditional (local) independence, 38–39
- Structural equation model, 588
- Target metric, 448
- Target total information function, 35
- Test characteristic curve. *See Total characteristic curve (TCC)*
- Test characteristic curve equating. *See Total characteristic function equating*
- Testlet model, 189, 229
- Three-level model analysis. *See also Three-parameter logistic (3PL) model*
- overview, 574, 577–579, 587
 - three-level model analysis–nutrition literacy, 1me4, 582–587, 583t–586t
 - three-level model analysis–nutrition literacy, proc glimmix, 579–582, 579t, 580t–581t
- Three-parameter logistic (3PL) model. *See also* Multidimensional three-parameter logistic (M3PL) model; Three-level model analysis
- application of the 3PL model to the mathematics data, MMLE, BILOG-MG, 192–195, 192f, 193t–194t, 195f
 - application of the 3PL model to the mathematics data, MMLE, mirt, 200–209, 201t–202t, 203f, 204t–208t, 209f, 210f, 211f, 212f
 - application of the 3PL model using MMLE and BILOG-MG, 192–195, 192t, 193t–194t, 195f, 196f
 - application of the 3PL model using MMLE and mirt, 200–209, 201t–202t, 203f, 204t–208t, 209f, 210f, 211f, 212f
 - appropriateness measurement and, 209, 211–216, 214t, 215f, 216f
 - calibration sample size and, 187–188
 - conceptual development of, 179–182, 181f
 - conceptual parameter estimation for, 183–186, 184f, 186f
 - conditional independence assessment and, 195–198, 196t, 197f
 - conditional independence assumption and, 188–192
 - information for, 216–220, 218f, 219f
 - issues to consider in selecting, 224–226
 - metric transformation and, 220, 221f
 - missing data and, 220–224
 - model–data fit assessment, 198–200, 199t
 - overview, 177–178, 179, 226–227
- Threshold, 128, 267, 268f
- Total characteristic curve (TCC), 114, 115f, 235, 236f, 254
- Total characteristic function, 111–114, 115f, 175–176
- Total characteristic function equating
- application of the total characteristic function equating method, EQUATE, 454–461, 456t–459t, 460t, 461f, 462t–463t, 464f, 465f
 - application of the total characteristic function equating method, SNSequate, 463–465, 466t–467t
- overview, 114, 450–454, 451f, 452f, 454f

- Total information area
 application of the NR model to a general science test, MMLE, *mirt*, 379, 379f
 application of the PC Model to a reasoning ability instrument, MMLE, *flexMIRT*, 255, 255f
 overview, 162–165, 164f, 166f, 177f
 three-parameter logistic (3PL) model and, 216–220, 218f, 219f
- Total multidimensional information, 407
- Total test information function, 177, 178f
- Trace line, 14–15, 14f, 17. *See also* Option response function (ORF)
- Transformation, metric. *See* Metric transformation
- Transformations. *See also* Metric transformation
 equating and, 446–454, 448f, 451f, 452f, 454f, 474
 nominal response (NR) model and, 359–360
 overview, 476–477
- Transition location parameter, 240–244
- Trichotomous model approach, 254–255
- True score theory, 5–6. *See also* Classical test theory (CTT)
- Truncated 2PL model, 186. *See also* Two-parameter logistic (2PL) model
- TSW likelihood ratio test, 486–487, 488–490
- Two-parameter logistic (2PL) model. *See also* Multidimensional two-parameter logistic (M2PL) model
 application of the 2PL model to the mathematics data, MMLE, BILOG-MG, 143–146, 143t, 145t–146t, 147t–148t, 148f
 application of the 2PL model to the mathematics data, MMLE, *mirt*, 152–162, 153t–158t, 160f, 161f
 application of the 2PL model using MMLE and BILOG-MG, 143–146, 143t, 145t–146t, 147t–148t
- application of the 2PL model using MMLE and *mirt*, 152–162, 153t–158t, 160f, 161f
- conceptual development of, 135–137, 136f
- conceptual parameter estimation for, 139–140, 140f
- fit assessment for, 146, 148–151, 148f, 151f
- generalized partial credit (GPC) model and, 351
- graded response (GR) model and, 352–353
- information and relative efficiency, 162–165, 164f, 166f
- information for, 137–139, 138f
- issues to consider in selecting, 224–226
- metric transformation and, 142–143
- nominal response (NR) model and, 365
- overview, 135, 136–137, 136f, 165, 167, 177–178
- Type I error rates, 190–191
- Type I Model C. *See* Bock–Samejima model for multiple-choice items (BS)
- Unconditional maximum likelihood estimation (UCON), 43–44. *See also* Likelihood function; Parameter estimates
- Unidimensionality assumption, 10–11, 21, 47, 171
- Uniform DIF, 480, 518–519. *See also* Differential item functioning (DIF) analysis
- Validity, 3, 75–76. *See also* Measurement
- Variable Map, 74, 75f
- Vector graphs, 401–404, 402f
- Vertical scaling, 475
- Visual analog scale, 354
- Weak principle of conditional (local) independence, 38–39
- Weighted MLE approach, 127

About the Author

R. J. de Ayala is Professor of Quantitative, Qualitative, and Psychometric Methods and Director of the Institutional Research Master's Program in the College of Educational and Human Sciences at the University of Nebraska–Lincoln. His research interests include psychometrics, item response theory, computerized adaptive testing, applied statistics, and multilevel models. His work has appeared in *Applied Psychological Measurement*, *Applied Measurement in Education*, the *British Journal of Mathematical and Statistical Psychology*, *Educational and Psychological Measurement*, the *Journal of Applied Measurement*, and the *Journal of Educational Measurement*. He is a Fellow of the American Psychological Association's Division 5: Evaluation, Measurement, and Statistics and of the American Educational Research Association. He is a recipient of a Big 12 Faculty Fellowship and holds a Gallup Research Professorship at UNL.