Testlet Response Theory and Its Applications

The measurement models employed to score tests have been evolving over the past century from those that focus on the entire test (true score theory) to models that focus on individual test items (item response theory) to models that use small groups of items (testlets) as the fungible unit from which tests are constructed and scored (testlet response theory, or TRT).

In this book, the inventors of TRT trace the history of this evolution and explain the character of modern TRT. Written for researchers and professionals in statistics, psychometrics, and educational psychology, the first part of the book offers an accessible introduction to TRT and its applications. The rest of the book is a comprehensive, self-contained discussion of the model couched within a fully Bayesian framework. Its parameters are estimated using Markov chain Monte Carlo procedures, and the resulting posterior distributions of the parameters yield insights into score stability that were previously unsuspected. The authors received the National Council on Measurement in Education award for scientific contribution to a field of educational measurement for this work.

HOWARD WAINER is a Distinguished Research Scientist for the National Board of Medical Examiners and Adjunct Professor of Statistics at the Wharton School of the University of Pennsylvania. Among his many honors, he received the Educational Testing Services' Senior Scientist award and is a Fellow of the American Statistical Association. This is his fifteenth book.

ERIC T. BRADLOW is the K. P. Chao Professor; Professor of Marketing, Statistics, and Education; and Academic Director of the Wharton Small Business Development Center at the Wharton School of the University of Pennsylvania. Before joining the Wharton faculty, he worked in the Corporate Marketing and Business Research Division at the DuPont Corporation and in the Statistics and Psychometrics Research Group at the Educational Testing Service. Bradlow was recently named a Fellow of the American Statistical Association.

XIAOHUI WANG is an Assistant Professor in the Department of Statistics at the University of Virginia. She worked as a Principal Data Analyst for three years in the Division of Data Analysis and Research Technology at the Educational Testing Service. She has twice received the National Council on Measurement in Education Award for Scientific Contribution to a Field of Educational Measurement.

Testlet Response Theory and Its Applications

HOWARD WAINER
National Board of Medical Examiners

ERIC T. BRADLOW
University of Pennsylvania

XIAOHUI WANG
University of Virginia

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521862721

First published 2007

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Wainer, Howard.
Testlet response theory and its applications / Howard Wainer, Eric T.
Bradlow, Xiaohui Wang.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-521-86272-1 (hardback)
ISBN-10: 0-521-86272-8 (hardback)
ISBN-13: 978-0-521-68126-1 (pbk.)
ISBN-10: 0-521-68126-X (pbk.)
1. Examinations – Scoring. 2. Bayesian statistical decision theory.
I. Bradlow, Eric T. II. Wang, Xiaohui III. Title.
LB3060.77.W35 2007
371.26 – dc22 2006030231

ISBN 978-0-521-86272-1 hardback
ISBN 978-0-521-68126-1 paperback

To Linda, Laura, and Weiqiang

Contents

Preface

This book describes the outcome of a research program that began in the mid-1980s. In the more than two decades since, many debts, both intellectual and financial, have been incurred. We are delighted to have the opportunity to acknowledge the help and express our gratitude.

First, organizational thanks are due to the National Board of Medical Examiners and its president, Donald Melnick, who has supported this work for the past five years. Senior Vice President Ronald Nungester and Associate Vice President Brian Clauser's enthusiasm for this work is especially appreciated.

Second, our gratitude to the Educational Testing Service (ETS), which employed all of us during the critical period in which the Bayesian version of testlet response theory was being birthed, is equally sincere. Henry Braun, then Vice President of Research at ETS, was a crucial voice in committees that provided support for this project. His enthusiasm and wise counsel were appreciated then and now. The funding organizations that provided support were the Graduate Record Board, the TOEFL Board, the College Board, the Joint Scientific Research and Development Committee, the Law School Admissions Council, and the research budget of ETS. We are also grateful to Kurt Landgraf, president of ETS, whose wise leadership made continued support of basic research financially possible.

Intellectual debts are harder to keep track of than financial ones, and if we have omitted anyone we hope that the lapse will be recognized as one of memory and not ingratitude. Our thanks, in more-or-less chronological order, go to

Gerald Kiely, who spent a summer at ETS as an intern and helped with the birth of the testlet concept.

Paul Rosenbaum, who before he went on to fame as the world's leading expert on observational studies occupied many hours discussing the use of item bundles as a viable practical solution to various pressing problems in modern testing.

His proof of the plausible conditional independence of bundles provided important supporting evidence for the concept.

David Thissen, whose collaboration on the use of a polytomous IRT model to score testlet-based tests was key to making the ideas easily operational and who helped us show how its use affects reliability.

Charles Lewis, for his penetrating analyses of test construction and scoring as well as his gentle insistence on continued exploration to make whatever we had better.

Stephen Sireci, who as a summer intern was the operational force behind the development of the initial testlet methods for studying DIF and for proving the importance of the testlet concept by showing the extent to which ignoring local dependence upwardly biases estimates of test reliability.

Robert Mislevy, who pointed out the critical importance of Robert Gibbons' bi-factor test scoring model as a more general testlet model, and who has made himself available countless times for discussions of issues both technical and logical; most recently on some vexing questions on the identifiability of parameters when covariates are used.

Zuru Du, another ETS summer intern, who came to Princeton in search of a thesis topic and left with the task of expanding the 2-PL testlet model to accommodate guessing. His success at doing this earned him his PhD and subsequent awards, as well as our thanks.

Cees Glas, who demonstrated that maximum likelihood estimation is not yet obsolete by using it to fit the 3-PL testlet model. He also wrote the code that did it by lunchtime.

The enjoyment, as well as the value, of writing a book loses more than just its patina of scholarly accomplishment if there is no prospect of it ever being published. Thus, our gratitude to Cambridge University Press and mathematics editor Lauren Cowles for ready acceptance and encouragement is much more than pro forma. When we first discussed the pros and cons of Cambridge doing the book Lauren pointed out that they were the outlet Newton chose to publish his *Principia.* That was enough of a recommendation for us.

Any project whose execution spanned decades accumulates debts to many others whose help was important. Most important are Elizabeth Brophy, Editha Chase, and Martha Thompson, whose work ethic and keen sense of organization kept everything in order.

Our gratitude to Navdeep Singh and the rest of his production staff for their help in making this book the best it could be.

The writing of this book was guided, in part, by the wisdom of two great physicists – Albert Einstein, who pointed out that "everything should be as

simple as possible; but no simpler," and J. Robert Oppenheimer, who in his class on quantum mechanics told his students that "I may be able to make it clearer, but I can't make it simpler."

The book begins conceptually, requiring no more mathematics than simple algebra and some basic statistical concepts, but it ends up much more technical. Part I (Chapters 1–6) is no more complex than any modern book on test theory, although there is an occasional integral.

Part II (Chapters 7–12), where we introduce and develop Bayesian testlet response theory, requires more. Here the technology of Bayesian methods with its requirement of distribution theory and an understanding of how a Markov process can eventually converge to a stationary result is an area that is relatively new and may require slower going. We have included a tutorial in these methods at the very end of the book (Chapter 15) that can be helpful in two ways. It provides an introduction that may be helpful by itself – its glossary can aid those for whom some of the language is unfamiliar – and its reference list can point interested readers to further details.

Part III (Chapters 13–15) contains two applications of the new model and its associated technology. These can be read without a full understanding of the mathematics and hopefully will whet the reader's appetite enough to warrant gaining the technical expertise for a deeper understanding.

Last, we have prepared a computer program (SCORIGHT) that can do everything we describe here. It is available free from us. You can download SCORIGHT's users manual from www.cambridge.org/9780521681261 as well as a permission-to-use form. To get a copy of the current version of SCORIGHT merely print out the permission form, sign it, and mail it (with your email address) to:

Howard Wainer
Distinguished Research Scientist
National Board of Medical Examiners
3750 Market Street
Philadelphia, Pennsylvania 19104