ITEM RESPONSE THEORY FOR PSYCHOLOGISTS

SUSAN E. EMBRETSON
STEVEN P. REISE

Item Response Theory for Psychologists

MULTIVARIATE APPLICATIONS BOOKS SERIES

The Multivariate Applications book series was developed to encourage the use of rigorous methodology in the study of meaningful scientific issues, and to describe the applications in easy to understand language. The series is sponsored by the Society of Multivariate Experimental Psychology and welcomes methodological applications from a variety of disciplines, such as psychology, public health, sociology, education, and business. Books can be single authored, multiple authored, or edited volumes. The ideal book for this series would take on one of several approaches: (1) demonstrate the application of a variety of multivariate methods to a single, major area of research; (2) describe a methodological procedure or framework that could be applied to a variety of research areas; or (3) present a variety of perspectives on a controversial topic of interest to applied researchers.

There are currently four books in the series:

1. *What if There Were No Significance Tests?*, co-edited by Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger (1997).
2. *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications and Programming*, written by Barbara M. Byrne (1998).
3. *Multivariate Applications in Substance Use Research*, co-edited by Jennifer S. Rose, Laurie Chassin, Clark C. Presson, and Steven J. Sherman (2000).
4. *Item Response Theory for Psychologists*, co-authored by Susan E. Embretson and Steven P. Reise.

Interested persons should contact the editor, Lisa L. Harlow, at: Department of Psychology, University of Rhode Island, 10 Chafee Rd., Suite 8, Kingston, RI 02881-0808; Phone: 401-874-4242; FAX: 401-874-5562; or E-Mail: LHarlow@uri.edu. Information can also be obtained from one of the editorial board members: Leona Aiken (Arizona State University), Gwyneth Boodoo (Educational Testing Service), Barbara Byrne (University of Ottawa), Scott Maxwell (University of Notre Dame), David Rindskopf (City University of New York), or Steve West (Arizona State University).

Item Response Theory for Psychologists

Susan E. Embretson
University of Kansas

Steven P. Reise
University of California


LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2000 Mahwah, New Jersey London

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430-2262

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Embretson, Susan E.
Item response theory for psychologists / Susan E. Embretson and Steven P. Reise.
p. cm. — (Multivariate applications)
Includes bibliographical references and index.
ISBN 0-8058-2818-4 (cloth : alk. paper) — ISBN 0-8058-2819-2 (pbk. : alk. paper)
1. Item response theory. 2. Psychometrics. I. Reise, Steven P. II. Title. III. Multivariate applications book series.
BF39.E495 2000
150'.28'7—dc21 99-048454
CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability

Printed in the United States of America
10 9 8 7 6

To Marshall and to the many IRT scholars
who have shaped the field and taught us their insights.
—Susan

To my parents, Ben and Ruth,
who provided support and inspiration throughout my pursuit
of higher education.
—Steve

Contents

Preface ix

PART I: INTRODUCTION

Chapter 1 Introduction 3

PART II: ITEM RESPONSE THEORY PRINCIPLES: SOME CONTRASTS AND COMPARISONS

Chapter 2 The New Rules of Measurement 13

Chapter 3 Item Response Theory as Model-Based Measurement 40

PART III: THE FUNDAMENTALS OF ITEM RESPONSE THEORY

Chapter 4 Binary IRT Models 65

Chapter 5 Polytomous IRT Models 95

Chapter 6	The Trait Level Measurement Scale: Meaning, Interpretations, and Measurement-Scale Properties	**125**
Chapter 7	Measuring Persons: Scoring Examinees with IRT Models	**158**
Chapter 8	Calibrating Items: Estimation	**187**
Chapter 9	Assessing the Fit of IRT Models	**226**

PART IV: APPLICATIONS OF IRT MODELS

Chapter 10	IRT Applications: DIF, CAT, and Scale Analysis	**249**
Chapter 11	IRT Applications in Cognitive and Developmental Assessment	**273**
Chapter 12	Applications of IRT in Personality and Attitude Assessment	**306**
Chapter 13	Computer Programs for Conducting IRT Parameter Estimation	**326**
	References	**345**
	Author Index	**363**
	Subject Index	**368**

Preface

The purpose of this book is to explain the new measurement theory to a primarily psychological audience. Item response theory (IRT) is not only the psychometric theory underlying many major tests today, but it has many important research applications. Unfortunately, the few available textbooks are not easily accessible to the audience of psychological researchers and practitioners; the books contain too many equations and derivations and too few familiar concepts. Furthermore, most IRT texts are slanted toward understanding IRT application within the context of large-scale educational assessments, such as analyzing the SAT. Our approach is more geared toward a psychological audience that is familiar with small-scale cognitive and personality measures or that wants to use IRT to analyze scales used in their own research.

Herein, familiar psychological concepts, issues, and examples are used to help explain various principles in IRT. We first seek to develop the reader's intuitive understanding of IRT principles by using graphical displays and analogies to classical measurement theory. Then, the book surveys contemporary IRT models, estimation methods, and computer programs. Because many psychological tests use rating scales, polytomous IRT models are given central coverage. Applications to substantive research problems, as well as to applied testing issues, are described.

The book is intended for psychology professionals and graduate students who are familiar with testing principles and classical test theory (CTT), such as covered in a graduate textbook on psychological testing (e.g., Anastasi & Urbina, 1997). Furthermore, the reader should have had a first-year sequence in graduate statistics, such as required in most psychology graduate programs. The reader need not have further training in either

statistics or measurement, however, to read this book. Although equations are necessary to present IRT models and estimation methods, we attempt to define all symbols thoroughly and explain the equations verbally or graphically.

The book is appropriate as a graduate textbook for a measurement course; in fact, drafts of the text have been used at the University of Kansas, University of California, Los Angeles, and the University of Virginia. We wish to thank students in these courses for finding numerous typos and for their comments on principles that have helped us improve the treatment of several topics. Although the book is most appropriate for psychology measurement courses, nothing precludes use in related fields. In fact, it can be used in schools of education, as well as in other social sciences and related areas, such as behavioral medicine and gerontology, where it might be used to explain measurement principles.

The idea for this book emerged during an American Psychological Association meeting in 1996. Susan Embretson had presented a paper on IRT in a session entitled "What Every Psychologist Should Know About Measurement—but Doesn't." In this paper, a foundation for chapter 2 in this book, the lack of an appropriate textbook was cited as one reason why psychologists are largely unfamiliar with IRT. Unintentionally, Susan convinced herself to write it—that is, if she could find the right coauthor. But who? Well, she thought, if Steve Reise were willing, maybe this textbook could be written. A few hours later, Steve appeared at an APA reception. To the point, Susan asked immediately "How about writing a book on IRT for psychologists with me?" "Yeah, good idea," replied Steve in his characteristic low-key manner.

Of course, the path from ideas to products is long. Although the writing was easy enough, many inconsistencies and incomparabilities in the literature and in the computer programs for estimating IRT parameters created difficulty. Developing a clear exposition requires unified concepts and generalities. The book was not finished until Spring 1999.

We think that two important communalities influenced the book. First, we have both taught measurement and IRT in departments of psychology. Thus, we are familiar with the conceptual issues that develop in teaching the psychology audience. Second, we are both Psychology PhDs of the University of Minnesota. Susan finished in 1973, and Steve in 1990. Thus, we share some perspectives on the role of measurement in psychology. However, we also differ in several ways. Susan struggled with IRT principles during the early years of its development. She was interested then, and now, in developing IRT models and methodology to interface cognitive theory and psychometrics. She focused primarily on Rasch-family models for binary data. Steve, in contrast, is a more recent PhD. He was interested in interfacing IRT models with personality measurement. He has concentrated primarily on complex IRT models (i.e., those with discrimination pa-

rameters) for rating scale data. These intellectual differences have enabled the book to include greater bredth of coverage and to elaborate differing perspectives on IRT models.

ACKNOWLEDGMENTS

We have relied on many good colleagues for their critiques of the chapters. IRT is not an easy field; the scholarship of our colleagues has been invaluable. We wish to thank the following persons for reading one or more chapters: Terry Ackerman, R. Darrell Bock, Paul DeBoeck, Fritz Drasgow, Niahua Duan, Mathilde Dutoit, Jan-Eric Gustafsson, Mark Haviland, Karen McCollam, Robert Mislevy, Eiji Muraki, Michael Nering, Mark Reckase, Lynne Steinberg, Jurgen Rost, David Thissen, Niels Waller, and Michael Yoes. Of course, these readers are not responsible for any remaining problems in the book. The book departs in many ways from typical treatments; IRT principles are explained in a simple and direct way. However, accomplishing simplicity sometimes requires obscuring issues that some scholars find important. We are very interested in improving the book in later revisions, so we urge the readers to address any comments or problems to us directly.

Last, but not least, we would like to thank those persons close to us. Marshall Picow, Susan's husband, knows what effort it has required to finish this book. He has been most helpful and patient while Susan has chained herself to the computer for days on end. He deserves much thanks for his continuing support. Steve thanks several scholars who have been of tremendous assistance throughout his career as an IRT researcher: David J. Weiss, Auke Tellegen, and his research associates Niels G. Waller and Keith Widaman.

—Susan E. Embretson
—Steven P. Reise

I

INTRODUCTION

1

Introduction

In an ever-changing world, psychological testing remains the flagship of applied psychology. Although the specific applications and the legal guidelines for using tests have changed, psychological tests have been relatively stable. Many well-known tests, in somewhat revised forms, remain current. Furthermore, although several new tests have been developed in response to contemporary needs in applied psychology, the principles underlying test development have remained constant. Or have they?

In fact, the psychometric basis of tests has changed dramatically. Although classical test theory (CTT) has served test development well over several decades, item response theory (IRT) has rapidly become mainstream as the theoretical basis for measurement. Increasingly, standardized tests are developed from IRT due to the more theoretically justifiable measurement principles and the greater potential to solve practical measurement problems.

This chapter provides a context for IRT principles. The current scope of IRT applications is considered. Then a brief history of IRT is given and its relationship to psychology is discussed. Finally, the purpose of the various sections of the book is described.

SCOPE OF IRT APPLICATIONS

IRT now underlies several major tests. Computerized adaptive testing, in particular, relies on IRT. In computerized adaptive testing, examinees receive items that are optimally selected to measure their potential. Differ-

ent examinees may receive no common items. IRT principles are involved in both selecting the most appropriate items for an examinee and equating scores across different subsets of items. For example, the Armed Services Vocational Aptitude Battery, the Scholastic Aptitude Test (SAT), and the Graduate Record Examination (GRE) apply IRT to estimate abilities. IRT has also been applied to several individual intelligence tests, including the Differential Ability Scales, the Woodcock-Johnson Psycho-Educational Battery, and the current version of the Stanford-Binet, as well as many smaller volume tests. Furthermore, IRT has been applied to personality trait measurements (see Reise & Waller, 1990), as well as to attitude measurements and behavioral ratings (see Engelhard & Wilson, 1996). Journals such as *Psychological Assessment* now feature applications of IRT to clinical testing issues (e.g., Santor, Ramsey, & Zuroff, 1994).

Many diverse IRT models are now available for application to a wide range of psychological areas. Although early IRT models emphasized dichotomous item formats (e.g., the Rasch model and the three-parameter logistic model), extensions to other item formats has enabled applications in many areas; that is, IRT models have been developed for rating scales (Andrich, 1978b), partial credit scoring (Masters, 1982), and multiple category scoring (Thissen & Steinberg, 1984). Effective computer programs for applying these extended models, such as RUMM, MULTILOG, and PARSCALE, are now available (see chap. 13 for details). Thus, IRT models may now be applied to measure personality traits, moods, behavioral dispositions, situational evaluations, and attitudes as well as cognitive traits.

The early IRT applications involved primarily unidimensional IRT models. However, several multidimensional IRT models have been developed. These models permit traits to be measured by comparisons within tests or within items. Bock, Gibbons, and Muraki (1988) developed a multidimensional IRT model that identifies the dimensions that are needed to fit test data, similar to an exploratory factor analysis. However, a set of confirmatory multidimensional IRT models have also been developed. For example, IRT models for traits that are specified in a design structure (like confirmatory factor analysis) have been developed (Adams, Wilson, & Wang, 1997; Embretson, 1991, 1997; DiBello, Stout, & Roussos, 1995). Thus, person measurements that reflect comparisons on subsets of items, change over time, or the effects of dynamic testing may be specified as the target traits to be measured. Some multidimensional IRT models have been closely connected with cognitive theory variables. For example, person differences in underlying processing components (Embretson, 1984; Whitely, 1980), developmental stages (Wilson, 1985) and qualitative differences between examinees, such as different processing strategies or knowledge structures (Kelderman & Rijkes, 1994; Rost, 1990) may be measured with the special IRT models. Because many of these models also

have been generalized to rating scales, applications to personality, attitude, and behavioral self-reports are possible, as well. Thus many measurement goals may be accommodated by the increasingly large family of IRT models.

HISTORY OF IRT

Two separate lines of development in IRT underlie current applications. In the United States, the beginning of IRT is often traced to Lord and Novick's (1968) classic textbook, *Statistical Theories of Mental Test Scores.* This textbook includes four chapters on IRT, written by Allan Birnbaum. Developments in the preceding decade provided the basis for IRT as described in Lord and Novick (1968). These developments include an important paper by Lord (1953) and three U.S. Air Force technical reports (Birnbaum, 1957, 1958a, 1958b). Although the air force technical reports were not widely read at the time, Birnbaum contributed the material from these reports in his chapters in Lord and Novick's (1968) book.

Lord and Novick's (1968) textbook was a milestone in psychometric methods for several reasons. First, these authors provided a rigorous and unified statistical treatment of test theory as compared to other textbooks. In many ways, Lord and Novick (1968) extended Gulliksen's exposition of CTT in *Theory of Mental Tests,* an earlier milestone in psychometrics. However, the extension to IRT, a much more statistical version of test theory, was very significant. Second, the textbook was well connected to testing. Fred Lord, the senior author, was a long-time employee of Educational Testing Service. ETS is responsible for many large-volume tests that have recurring psychometric issues that are readily handled by IRT. Furthermore, the large sample sizes available were especially amendable to statistical approaches. Third, the textbook was well connected to leading and emerging scholars in psychometric methods. Lord and Novick (1968) mentioned an ongoing seminar at ETS that included Allan Birnbaum, Michael W. Browne, Karl Joreskog, Walter Kristof, Michael Levine, William Meredith, Samuel Messick, Roderick McDonald, Melvin Novick, Fumiko Samejima, J. Philip Sutcliffe, and Joseph L. Zinnes in addition to Frederick Lord. These individuals subsequently became well known for their contributions to psychometric methods.

R. Darrell Bock, then at the University of North Carolina, was inspired by the early IRT models, especially those by Samejima. Bock was interested in developing effective algorithms for estimating the parameters of IRT models. Subsequently, Bock and several student collaborators at the University of Chicago, including David Thissen, Eiji Muraki, Richard Gibbons, and Robert Mislevy, developed effective estimation methods

and computer programs, such as BILOG, TESTFACT, MULTILOG, and PARSCALE. In conjunction with Murray Aitken (Bock & Aitken, 1981), Bock developed the marginal maximum likelihood method to estimate the parameters, which is now considered state of the art in IRT estimation. An interesting history of IRT, and its historical precursors, was published recently by Bock (1997).

A rather separate line of development in IRT may be traced to Georg Rasch (1960), a Danish mathematician who worked for many years in consulting and teaching statistics. He developed a family of IRT models that were applied to develop measures of reading and to develop tests for use in the Danish military. Rasch (1960) was particularly interested in the scientific properties of measurement models. He noted that person and item parameters were fully separable in his models, a property he elaborated as *specific objectivity*. Andersen (1972), a student of Rasch, consequently elaborated effective estimation methods for the person and item parameters in Rasch's models.

Rasch inspired two other psychometricians who extended his models and taught basic measurement principles. In Europe, Gerhard Fischer (1973) from the University of Vienna, extended the Rasch model for binary data so that it could incorporate psychological considerations into the parameters. Thus stimulus properties of items, treatment conditions given to subjects, and many other variables could be used to define parameters in the linear logistic latent trait model. This model inspired numerous applications and developments throughout Europe. Fischer's (1974) textbook on IRT was influential in Europe but had a restricted scope since it was written in German.

Rasch visited the United States and inspired Benjamin Wright, an American psychometrician, to subsequently teach objective measurement principles and to extend his models. Rasch visited the University of Chicago, where Wright was a professor in education, to give a series of lectures. Wright was particularly inspired by the promise of objective measurement. Subsequently, a large number of doctoral dissertations were devoted to the Rasch model under Wright's direction. Several of these PhDs became known subsequently for their theoretical contributions to Rasch-family models, including David Andrich (1978a), Geoffrey Masters (1982), Graham Douglas (Wright & Douglas, 1977), and Mark Wilson (1989). Many of Wright's students pioneered extended applications in educational assessment and in behavioral medicine. Wright also lectured widely on objective measurement principles and inspired an early testing application by Richard Woodcock in the Woodcock-Johnson Psycho-Educational Battery.

Rather noticeable by its absence, however, is the impact of IRT on psychology. Wright's students, as education PhDs, were employed in

education or in applied settings rather than in psychology. Bock's affliation at the University of Chicago also was not primarily psychology, and his students were employed in several areas but rarely psychology.

Instead, a few small pockets of intellectual activity could be found in psychology departments with programs in quantitative methods or psychometrics. The authors are particularly familiar with the impact of IRT on psychology at the University of Minnesota, but similar impact on psychology probably occurred elsewhere. Minnesota had a long history of applied psychological measurement. In the late 1960s and early 1970s, two professors at Minnesota—Rene Dawis and David Weiss—became interested in IRT. Dawis was interested in the objective measurement properties of the Rasch model. Dawis obtained an early version of Wright's computer program through Richard Woodcock, who was applying the Rasch model to his tests. Graduate students such as Merle Ace, Howard Tinsley, and Susan Embretson published early articles on objective measurement properties (Tinsley, 1972; Whitely[1] & Dawis, 1976). Weiss, on the other hand, was interested in developing computerized adaptive tests and the role for complex IRT models to solve the item selection and test equating problems. Graduate students who were involved in this effort included Isaac Bejar, Brad Sympson, and James McBride. Later students of Weiss, including Steve Reise, moved to substantive applications such as personality.

The University of Minnesota PhDs had significant impact on testing subsequently, but their impact on psychological measurement was limited. Probably like other graduate programs in psychology, new PhDs with expertise in IRT were actively recruited by test publishers and the military testing laboratories to implement IRT in large volume tests. Although this career path for the typical IRT student was beneficial to testing, psychology remained basically unaware of the new psychometrics. Although (classical) test theory is routine in the curriculum for applied psychologists and for many theoretically inclined psychologists, IRT has rarely had much coverage. In fact, in the 1970s and 1980s, many psychologists who taught measurement and testing had little or no knowledge of IRT. Thus the teaching of psychological measurement principles became increasingly removed from the psychometric basis of tests.

THE ORGANIZATION OF THIS BOOK

As noted in the brief history given earlier, few psychologists are well acquainted with the principles of IRT. Thus most psychologists' knowledge of the "rules of measurement" is based on CTT. Unfortunately, under IRT many well-known rules of measurement derived from CTT no longer ap-

[1]Susan E. Embretson has also published as Susan E. Whitely.

ply. In fact, some new rules of measurement conflict directly with the old rules. IRT is based on fundamentally different principles than CTT. That is, IRT is model-based measurement that controls various confounding factors in score comparisons by a more complete parameterization of the measurement situation.

The two chapters in Part II, "Item Response Theory Principles: Some Contrasts and Comparisons," were written to acquaint the reader with the differences between CTT and IRT. Chapter 2, "The New Rules of Measurement," contrasts 10 principles of CTT that conflict with corresponding principles of IRT. IRT is not a mere refinement of CTT; it is a different foundation for testing. Chapter 3, "Item Response Theory as Model-Based Measurement," presents some reasons why IRT differs fundamentally from CTT. The meaning and functions of measurement models in testing are considered, and a quick overview of estimation in IRT versus CTT is provided. These two chapters, taken together, are designed to provide a quick introduction and an intuitive understanding of IRT principles that many students find difficult.

More extended coverage of IRT models and their estimation is included in Part III, "The Fundamentals of Item Response Theory." Chapter 4, "Binary IRT Models," includes a diverse array of models that are appropriate for dichotomous responses, such as "pass versus fail" and "agree versus disagree." Chapter 5, "Polytomous IRT Models," is devoted to an array of models that are appropriate for rating scales and other items that yield responses in discrete categories. Chapter 6, "The Trait Level Scale: Meaning, Interpretations and Measurement Scale Properties," includes material on the various types of trait level scores that may be obtained from IRT scaling of persons. Also, the meaning of measurement scale level and its relationship to IRT is considered. Chapters 7 and 8, "Measuring Persons: Scoring Examinees with IRT Models" and "Calibrating Items: Estimation," concern procedures involved in obtaining IRT parameter estimates. These procedures differ qualitatively from CTT procedures. The last chapter in this section, "Assessing the Fit of IRT Models" (chap. 9), considers how to decide if a particular IRT model is appropriate for test data.

The last section of the book, "Applications of IRT Models," is intended to provide examples to help guide the reader's own applications. Chapter 10, "IRT Applications: DIF, CAT, and Scale Analysis," concerns how IRT is applied to solve practical testing problems. Chapters 11 and 12, "IRT Applications in Cognitive and Developmental Assessment" and "IRT Applications in Personality and Attitude Assessment," consider how IRT can contribute to substantive issues in measurement. The last chapter of the book, "Computer Programs for IRT Models," gives extended coverage to the required input and the results produced from several selected computer programs.

Although one more chapter originally was planned for the book, we decided not to write it. IRT is now a mainstream psychometric method, and the field is expanding quite rapidly. Our main concern was to acquaint the reader with basic IRT principles rather than to evaluate the current state of knowledge in IRT. Many recurring and emerging issues in IRT are mentioned throughout the book. Perhaps a later edition of this book can include a chapter on the current state and future directions in IRT. For now, we invite readers to explore their own applications and to research issues in IRT that intrigue them.

References

Ackerman, T. A. (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. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18,* 257-275.

Adams, R. A., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), Objective measurement III: Theory into practice. Norwood, NJ: Ablex.

Adams, R. A., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

American College Testing. (1993). *COMPASS user's guide.* Iowa City, IA.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice-Hall.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *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,* 42-54.

Andersen, E. B. (1995). Polytomous Rasch models and their estimation. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Andrich, D. (1978a). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR 20 index and the Guttman scale response pattern. *Educational Research and Perspectives, 9,* 95-104.

Andrich, D. (1988a). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1,* 363-378.

Andrich, D. (1988b). *Rasch models for measurement.* Newbury Park, CA: Sage.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19,* 101–119.

Andrich, D. (1996). Theoretical and empirical evidence on the dichotomization of graded responses. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice.* Norwood, NJ: Ablex.

Andrich, D. (1997). An hyperbolic cosine IRT model for unfolding direct response of persons to items. In W. J. Van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer-Verlag.

Andrich, D., & Styles, I. (1994). Psychometric evidence of growth spurts in early adolescence. *Journal of Early Adolescence, 14,* 328–344.

Angoff, W. (1982). Summary and derivation of equating methods used at ETS. In P. Holland & D. Rubin (Eds.), *Test equating.* New York: Academic Press.

Assessment Systems Corporation. (1991). *MicroCAT 3.0* [computer program]. St. Paul, MN: Assessment Systems Corporation.

Assessment Systems Corporation. (1996). *User's manual for the XCALIBRE marginal maximum-likelihood estimation program.* St. Paul, MN: Assessment Systems Corp.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17,* 239–251.

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.

Balasubramanian, S. K., & Kamakura, W. A. (1989). Measuring consumer attitudes toward the marketplace with tailored interviews. *Journal of Marketing Research, 26,* 311–326.

Baumeister, R. F., & Tice, D. M. (1988). Metatraits. *Journal of Personality, 56,* 571–598.

Bejar, I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English (College Board Report No. 81-8).* Princeton, NJ: Educational Testing Service, 1981. (ETS No. 81-35).

Bentler, P. M., & Wu, E. J. C. (1995). *EQS for Windows User's Guide.* Encino, CA: Mulivariate Software, Inc.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Annee Psychologique, 11,* 245–336.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10,* 167–174.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems.* Series Report No. 58-16. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas: January.

Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability.* Technical Report No. 17. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Birnbaum, A. (1958b). *On the estimation of mental ability.* Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas: January.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16,* 21–33.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Bock, R. D., Thissen, D., & Zimowsky, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34,* 197–211.

Bock, R. D., & Zimowski, M. F. (1996). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Brainerd, C. J. (1978). The stage question in cognitive-developmental theory. *The Behavioral and Brain Sciences, 2,* 173–213.

Buck, G., Tatsuoka, K., & Kostin, I. (in press). Exploratory rule space analysis of the Test of English for International Communication. *Journal of Language and Teaching.*

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97.*

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10,* 1–19.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Chopin, B. (1968). An item bank using sample-free calibration. *Nature, 219,* 870–872.

Chopin, B. (1983). *A fully conditional estimation procedure for Rasch model parameters.* Report No. 196. Los Angeles, CA: University of California, Graduate School of Education Center for the Study of Evaluation.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–32.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

College Board. (1990). *Coordinator's notebook for the computerized placement tests.* Princeton, NJ: Educational Testing Service.

Collins, L., & Horn, J. (1991). *Best methods for analyzing change* (pp. 184–197). Washington, DC: American Psychological Association Books.

Cook, W. W., & Medley, D. M. (1954). Proposed hostility and pharisaic-virtue scales for the MMPI. *Journal of Applied Psychology, 38,* 414–418.

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1992). *The revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Brown (Eds.), *Test validity*. Hillsdale, NJ: Lawrece Erlbaum Associates.

Cronbach, L., & Furby, L. (1970). How should we measure change—Or should we? *Psychological Bulletin, 74,* 68–80.

Cronbach, L., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley and Sons.

Curran, L. T., & Wise, L. L. (1994). *Evaluation and implementation of CAT-ASVAB.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on DAS and KAIT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regressions. *Psychological Bulletin, 107,* 394–400.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16,* 327–343.

DeLeeuw, J., & Verhelst, N. D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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). *Determining the sensitivity of CAT-ASVAB scores to changes in item response curves with medium of administration* (Report No. 86-189). Alexandria, VA: Center for Naval Analyses.

Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves* (Report No. 86-24). Alexandria, VA: Center for Naval Analyses.

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. Wilson (Ed.), *Objective measurement: Theory and practice* (Vol. 2, pp. 201–317). Norwood, NJ: Ablex.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Methods, 19,* 5–22.

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.

Douglas, J., Roussos, L. A., & Stout, W. F. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33,* 465–485.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92,* 526–531.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Drasgow, F., & Parsons, C. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

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., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9,* 47–64.

Educational Testing Service. (1993). *GRE 1993–94 guide to the use of the Graduate Record Examinations Program.* Princeton, NJ: ETS.

Educational Testing Service. (1998). *Grow in school: Achievement gains from the fourth to the eight grade.* Policy Information Center: Princeton, NJ: ETS.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24,* 133–148.

Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. *International Journal of Psychology, 24,* 665–684.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E. (1988, June). *Psychometric models and cognitive design systems.* Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–516.

Embretson, S. E. (1994a). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1994b). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues.* Ottawa: Edumetric Research Group, University of Ottawa.

Embretson, S. E. (1995a). Developments toward a cognitive design system for psychological tests. In D. Lupinsky & R. Dawis (Eds.), *Assessing individual differences in human behavior.* Palo Alto, CA: Davies-Black Publishing Company.

Embretson, S. E. (1995b). A measurement model for linking individual change to processes and knowledge: Application to mathematical learning. *Journal of Educational Measurement, 32,* 277–294.

Embretson, S. E. (1995c, August). *The new rules of measurement.* Paper presented at the annual meeting of the American Psychological Association, New York.

Embretson, S. E. (1995d). The role of working memory capacity and general control processes in intelligence. *Intelligence, 20,* 169–190.

Embretson, S. E. (1995e, June). *Structured latent trait modes in measuring the modifiability of ability to stress.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

Embretson, S. E. (1996a). Item response theory models and inferential bias in multiple group comparisons. *Applied Psychological Measurement, 20,* 201–212.

Embretson, S. E. (1996b). The new rules of measurement. *Psychological Assessment, 8,* 341–349.

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective Measurement III* (pp. 223–236). Norwood, NJ: Ablex.

Embretson, S. E. (1998a). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 300–326.

Embretson, S. E. (1998b, August). *Modifiability in lifespan development: Multidimensional Rasch Model for learning and change.* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Embretson, S. E. (1998c, October). *Multidimensional measurement from dynamic tests: Abstract reasoning under stress.* Presidential address for the Society of Multivariate Experimental Psychology, Mahwah, NJ.

Embretson, S. E. (in press). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E., & McCollam, K. M. (in press). A multicomponent Rasch model for measuring covert processes. In M. Wilson & G. Engelhard. *Objective Measurement V.* In G. Engelhard & M. Wilson (Eds.), Objective measurement III: Theory into practice. Norwood, NJ: Ablex.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175–193.

Engelhard, G., & Wilson, M. (1996). (Eds.). *Objective measurement III: Theory into practice.* Norwood, NJ: Ablex.

Etazadi-Amoli, J., & McDonald, R. P. (1983). A second generation nonlinear factor analysis. *Psychometrika, 48,* 315–342.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H. (1974). *Einfuhrung in die Theorie psychologischer Tests* (Introduction to mental test theory.) Berne: Huber.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46,* 59–77.

Fischer, G. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Fischer, K. W., Pipp, S. L., & Bullock, D. (1984). Detecting discontinuities in development: Methods and measurement. In R. N. Ende & R. Harmon (Eds.), *Continuities and discontinuities in development.* Norwood, NJ: Ablex.

Fitzpatrick, S. J., Choi, S. W., Chen, S., Hou, L., & Dodd, B. G. (1994). IRTINFO: A SAS macro program to compute item and test information. *Applied Psychological Measurement, 18,* 390.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality, 29,* 168–188.

Fraser, C. (1988). *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, N.S.W.: University of New England, Centre for Behavioral Studies.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review, 69,* 355–365.

Galton, F. (1883). *Inquiry into human faculty and its development.* London: Macmillan.

Gangestad, S., & Snyder, M. (1985). To carve nature at its joints: On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349.

Gibbons, R. D., Clark, D. C., Cavanaugh, S. V., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19,* 43–55.

Gibbons, R. D., & Hedeker, D. R. (1992). Full information item bi-factor analysis. *Psychometrika, 57,* 423–436.

Gierl, M. J., & Ackerman, T. (1996). XCALIBRE Marginal maximum-likelihood estimation program, Windows version 1.10. *Applied Psychological Measurement, 20,* 303–307.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234–246.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23,* 443–451.

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.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 337–385.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287–302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer Academic Publishers.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications.* Boston: Kluwer Academic Publishers.

Hamilton, J. C., & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology, 59,* 1301–1307.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20,* 191–206.

Harris, C. W. (Ed.). (1963). *Problems in measuring change.* Madison: University of Wisconsin Press.

Harvey, R. J., & Murry, W. D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait analyses. *Journal of Personality Assessment, 62,* 116–129.

Harvey, R. J., Murry, W. D., & Markham, S. E. (1994). Evaluation of three short-form versions of the Meyers-Briggs Type Indicator. *Journal of Personality Assessment, 63,* 181–184.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49–78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1–14.

Haviland, M. G., & Reise, S. P. (1996). Structure of the twenty item Toronto alexithymia scale. *Journal of Personality Assessment, 66,* 116–125.

Hendryx, M. S., Haviland, M. G., Gibbons, R. D., & Clark, D. C. (1992). An application of item response theory to alexithymia assessment among abstinent alcoholics. *Journal of Personality Assessment, 58,* 506–515.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement, 18,* 197–204.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15,* 153–169.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–602.

Holland, P., & Rubin, D. (1982). *Test equating.* New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369–380.

Hoskens, M., & DeBoeck, P. (1997). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32,* 261–277.

Huang, D. C., Church, T. A., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 28,* 197–218.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 6,* 818–825.

Jannerone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357–373.

Jansen, M. G. H., & Roshkam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51,* 69–72.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Janssen, R., DeBoeck, P., & Van der Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence, 22,* 291–310.

Jones, L. V. The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Joreskog, K., & Sorbom, D. (1996). *LISREL 8: User's Reference Guide.* Chicago: Scientific Software Int.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149–176.

Kemtes, K. A., & Kemper, S. (1997). Item response theory analysis of Sarason's Cognitive Interference questionnaire: A multiple group comparison of age. Presented at the *New Rules of Measurement Conference,* Continuing Education: Lawrence, KS.

Kim, S. (1997). BILOG 3 for windows: Item analysis and test scoring with binary logistic models. *Applied Psychological Measurement, 21,* 371–376.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 116–130.

King, D. W., King, L. A., Fairbank, J. A., & Schlenger, W. E. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment, 5,* 457–471.

Kingsbury, G. G., & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice, 12,* 21–27, 39.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241–261.

Knowles, E. S. (1988). Item context effects on personality scales: measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312–320.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7,* 15–32.

Koch, W. R., Dodd, B. G., & Fitzpatrick, S. J. (1990). Computerized adaptive measurement of attitudes. *Measurement and Evaluation in counseling and Development, 23,* 20–30.

Labouvie-Vief, G., & Gonda, J. N. (1976). Cognitive strategy training and intellectual performance in the elderly. *Journal of Gerontology, 31,* 327-332.

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model mispecificaiton, robustness, and parameter linking. *Applied Psychological Measurement, 12,* 365-376.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46,* 467-474.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161-176.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. L. (1992). Distinguishing between item response theory models. *Applied Psychological Measurement, 16,* 261-278.

Levine, M. V., & Rubin, D. B. (1979). Measuring appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367-386.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension (Technical Report no. 162).* Urban, IL: Center for the study of Reading, University of Illinois, 1980.

Liou, M. (1993). Exact person tests for assessing model-data fit in the Rasch model. *Applied Psychological Measurement, 17,* 187-195.

Liou, M., & Chang, C. H. (1992). Constructing the exact significance level for a person-fit statistic. *Psychometrika, 2,* 169-181.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

Lord, F. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin, 72,* 336-339.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. N., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1,* 1-27.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1,* 477-482.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16,* 33-40.

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.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

Maris, E. M. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G. N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement, 21,* 19-32.

Masters, G. N. (1985). A comparison of latent trait and latent class analyses of Likert-type data. *Psychometrika, 50,* 69-82.

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10,* 355-367.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529-544.

Masters, G. N., & Wright, B. D. (1996). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

Maxwell, S., & Delaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85–93.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18,* 245–256.

McBride, J. R. (1997). Technical perspectives. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computer Adaptive Testing*. Washington, DC: American Psychological Association.

McCollam, K. M. Schmidt (1997). *The modifiability of age differences in spatial visualization*. Unpublished doctoral dissertation, University of Kansas, Lawrence, Kansas.

McCollam, K. M. Schmidt (1998). Latent trait and latent class models. In G. M. Marcoulides (Ed.), *Modern methods for business research*. Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1962). Nonlinear factor analysis. *Psychometric Monograph* No. 15.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

McDonald, R. P. (1985). Comments of D. J. Bartholomew, Foundations of factor analysis: Some practical implications. *British Journal of Mathematical and Statistical Psychology, 38,* 134–137.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23–40.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49–57.

McKinley, R. L., & Way, W. D. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models. *TOEFL technical report TR-5,* Princeton, NJ: Educational Testing Service, February.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Medina-Diaz, M. (1993). Analysis of cognitive structure using the lienar logistic test model and quadratic assignment. *Applied Psychological Measurement, 17,* 117–130.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311–314.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9,* 3–8.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111–120.

Meijer, R. R., Muijtjens, A. M. M., & van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9,* 77–89.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21,* 321–226.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261–272.

Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. *Psychometrika, 61,* 629–645.

Mellenbergh, G. J. (1994a). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Mellenbergh, G. J. (1994b). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223–236.

Merideth, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21,* 315–330.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287–304.

Milsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. (1986). Bayesian modal estimation I item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. (1993). Foundations of a new test theory. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG-3; Item analysis and test scoring with binary logistic models* [Computer software]. Mooresville, IN: Scientific Software.

Mislevy, R., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215.

Mitchell, K. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery and their relationship to success in Army training.* Unpublished doctoral dissertation, Cornell University.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75–106.

Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9,* 27–45.

Moreno, K. E., & Segall, D. O. (1992). CAT-ASVAB precision. *Proceedings of the 34th annual conference of the Military Testing Association, 1,* 22–26.

Moreno, K. E., Wetzel, D. C., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8,* 155–163.

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. (1993a). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351–363.

Muraki, E. (1993b). *POLYFACT* [Computer program]. Princeton, NJ: Educational Testing Service.

Muraki, E. (1996). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153–164). New York: Springer-Verlag.

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks.* Chicago: Scientific Software Int.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73–90.

Muraki, E., & Engelhard, G. (1985). Full information item factor analysis: Application of EAP scores. *Applied Psychological Measurement, 9,* 417–430.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551–560.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthen, B. (1987). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model. Theoretical integration and user's guide.* Mooresville, IN: Scientific Software.

Nandakumar, R. (1993). Assessing essential dimensionality of real data. *Applied Psychological Measurement, 17,* 29-38.

Nandakumar, R. (1994). Assessing dimensionality of a set of items—Comparison of different approaches. *Journal of Educational Measurement, 31,* 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18,* 41-68.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, 99,* 166-180.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil test: When are they equivalent? *Applied Psychological Measurement, 22,* 71-83.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z 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. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16,* 237-248.

Ozer, D. J., & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology, 45,* 357-388.

Park, D., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163-173.

Pascual-Leone, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica, 32,* 301-345.

Pellegrino, J. W., Mumaw, R., & Shute, V. (1985). Analyses of spatial aptitude and expertise. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Pennings, A. H., & Hessels, M. G. P. (1996). The measurement of mental attentional capacity: A neo-Piagetian developmental study. *Intelligence, 23,* 59-78.

Peterson, N., Marco, G., & Steward, E. (1982). A test of the adequacy of linear score equating models. In P. Holland & D. Rubin (Eds.), *Test equating.* New York: Academic Press.

Ragosa, D., Brandt, D., & Zimowskyk, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726-748.

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.

Ramsey, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Glegvad (Ed.), *The Danish Yearbook of Philosophy* (pp. 58-94). Copenhagen: Munksgaard.

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. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in item response theory. *Applied Psychological Measurement, 14,* 127-137.

Reise, S. P. (1995). Scoring method and the detection of response aberrancy in a personality assessment context. *Applied Psychological Measurement, 19,* 213–229.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217–226.

Reise, S. P., & Flannery, Wm. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9,* 9–26.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data: The parameterization of the Multidimensional Personality Questionnaire. *Applied Psychological Measurement, 14,* 45–58.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143–151.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 352–566.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20,* 231–255.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. Van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer-Verlag.

Roskam, E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. Van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer-Verlag.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12,* 397–409.

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 polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75–92.

Rost, J., Carstensen, C., & Davier, von, M. (1996). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences.* New York: Waxmann Munster.

Rost, J., & Georg, W. (1991). Alternative Skalierungsmoglichkeiten zur klassischen Testtheorie am Beispiel der Skala "Jugendzentrismus." *Zerntral-Archiv-Information,* 28.

Rost, J., & von Davier, M. (1992). *MIRA: A PC-program for the mixed Rasch model.* Kiel, Federal Republic of Germany: IPN-Institute for science Education.

Roussos, L., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese, 16,* 170–233.

Safrit, M. J., Costa, M. G., & Cohen, A. S. (1989). Item response theory and the measurement of motor behavior. *Research Quarterly for Exercise and Sport, 60,* 325–335.

Salthouse, T. A., & Mitchell, D. R. D. (1990). Effects of age and naturally occurring experience on spatial visualization performance. *Developmental Psychology, 26,* 845–854.

Santor, D. A., Ramsay, J. O., &, Zuroff, D. C. (1998). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6,* 255–270.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory.* New York: Springer.

Schaie, K. W., & Willis, S. L. (1993). Age difference patterns of psychometric intelligence in adulthood: Generalizability within and across ability domains. *Psychology and Aging, 8,* 44–55.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354.

Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTT as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34,* 333–354.

Sheehan, K. M. (1998). *Understanding students' underlying strengths and weaknesses: A tree-based regression approach.* Technical Report. Princeton, NJ: ETS.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16,* 65–76.

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27,* 255–272.

Sheridan, B., Andrich, D., & Luo, G. (1996). *Welcome to RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models.* User's Guide.

Siegler, R. S. (1981). Developmental sequences within and between groups. *Monographs of the Society for Research in Child Development, 46,* No. 2.

Smith, L. L., & Reise, S. P. (in press). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75,* 1350–1362.

Smith, R. M. (1996). Item component equating. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice, Volume 3.* Norwood, NJ: Ablex.

Smith, R. M., & Kramer, G. A. (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement, 52,* 835–846.

Snyder, M. (1974). Self monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30,* 526–537.

Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology, 51,* 125–139.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. Embretson (Ed.), *Test design: New directions in psychology and psychometrics* (pp. 169–193). New York: Academic Press.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161–169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5,* 417–426.

Spray, J. A. (1997). Multiple-attempt, single-item response models. In W. J. Van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer-Verlag.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26,* 261–271.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika, 48,* 259–267.

Steinberg, L. (1994). Context and serial order effects in personality measurement: Limits on the generality of "measuring changes the measure." *Journal of Personality and Social Psychology, 66,* 341–349.

Steinberg, L., & Jorgensen, R. (1996). Assessing the MMPI-based Cook-Medley Hostility scale: The implications of dimensionality. *Journal of Personality and Social Psychology, 70,* 1281–1287.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81–97.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 221–263.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm (Research Report 93-2).* Princeton NJ: ETS.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

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 unidimensional assessment and ability estimation. *Psychometrika, 55,* 293–326.

Stout, W. F., & Roussos, L. A. (1995). *SIBTEST users manual* (2nd ed.) [Computer program manual]. Urbana-Champaign: University of Illinois, Department of Statistics.

Stout, W., Hsin-Hung, L. Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21,* 195–213.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349–364.

Swaminathan, H., & Rogers, J. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Sykes, R. C., & Ito, K. (1997). The effects of computer administration on scores and item parameter estimates of an IRT-based licensure examination. *Applied Psychological Measurement, 21,* 57–63.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 34–38.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12,* 55–73.

Tatsuoka, K. K. (1996). Use of generalized person-fit indices, zetas for statistical pattern classification. *Applied Measurement in Education, 9,* 65–75.

Tatsuoka, K. K., Solomonson, C., & Singley, K. (in press). The new SAT I mathematics profile. In G. Buck, D. Harnish, G. Boodoo & K. Tatsuoka (Eds.), *The new SAT.*

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215–231.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221–230.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Methods, 7,* 81–96.

Taylor, G. J., Bagby, R. M., & Parker, J. D. A (1992). The revised Toronto Alexithymia Scale: Some reliability, validity, and normative data. *Psychotherapy and Psychosomatics, 57,* 34–41.

Tellegen, A. (1982). *Brief manual for the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota, Minneapolis.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56,* 621–663.

Thayer, R. E. (1989). *The biopsychology of mood and arousal.* New York: Oxford University Press.

Theunissen, T. J. J. M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381–389.

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*: Chicago: Scientific Software Int.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar, (Eds.), *Test theory for a new generation of tests* (pp. 79–97). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501–519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104,* 385–395.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D., Steinberg, L., Pyszczynksi, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analyses. *Applied Psychological Measurement, 7,* 211–226.

Thissen, D., 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: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1992). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tinsley, H. E. A. (1972). *An investigation of the Rasch simple logistic model for tests of intelligence or attainment.* Doctoral dissertation, University of Minnesota. Ann Arbor, MI: University Microfilms, No. 72-14387.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin,* 103, 299–314.

Townsend, J. T., & Ashby, G. (1984). Measurement scale and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 133–344.

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory.* New York: Springer.

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. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–186). New York: Springer-Verlag.

von Davier, M. (1995). *WINMIRA: A Program System for Analyses with the Rasch Model, with the Latent Class Analysis and with the Mixed Rasch Model.* University of Kiel, Germany: Institute for Science Education.

von Davier, M., & Rost, J. (1996). Self monitoring—A class variable? In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences.* New York: Waxmann Munster.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339–368.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology, 57,* 1051–1058.

Waller, N. G., Tellegen, A, McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64,* 545–576.

Wang, W.-C., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between and within items. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice, Volume 4,* 139–156.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2,* 217–282.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774–789.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36,* 329–337.

White, N., & Cunningham, W. R. (1987). The age comparative construct validity of speeded cognitive factors. *Multivariate Behavioral Research, 22,* 249–265.

Whitely, S. E., & Schneider, L. M. (1981). Information structure on geometric analogies: A test theory approach. *Applied Psychological Measurement, 5,* 383–397.

Wiggins, J. S. (1972). *Personality and prediction: Principles of personality assessment.* Malabor, FL.: Robert E. Kriger Publishing.

Willis, S. L., & Nesselroade, C. S. (1990). Long-term effects of fluid ability training in old-old age. *Developmental Psychology, 26,* 905–910.

Wilson, M. (1985). Measuring stages of growth: A psychometric model of hierarchical development. Occasional paper No. 19. Hawthorn, Victoria: Australian Council for Educational Research.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309–325.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific Software.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education, 2,* 235–241.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review, Monograph Supplement, 3*(6).

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery Revised.* Allen, TX: DLM Teaching Resources.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281–294.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement.* Chicago: Mesa Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1995). *MATS: Multi Aspect Test Software.* Camberwell Victoria, Australia: Australian Council for Educational Research.

Yamamoto, K. (1987). *A model that combines IRT and latent class models.* Unpublished doctoral dissertation, University of Illinois, Champaign, IL.

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. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Zara, A. R. (1988). Introduction to item response theory and computerized adaptive testing as applied in licensure and certification testing. *National Clearinghouse of Examination Information newsletter, 6,* 11–17.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71–88.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293–308.