Listening. Learning. Leading.®

ETS R&D Scientific and Policy Contributions Series

ETS SPC-13-05

Item Response Theory

James E. Carlson

Matthias von Davier

December 2013

ETS R&D Scientific and Policy Contributions Series

SERIES EDITOR

Randy E. Bennett
Norman O. Frederiksen Chair in Assessment Innovation

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Managing Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Gary Ockey
Research Scientist

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Senior Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Director, Research

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ruth Greenwood
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS R&D Scientific and Policy Contributions series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS R&D Scientific and Policy Contributions series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of ETS.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Item Response Theory

James E. Carlson and Matthias von Davier
Educational Testing Service, Princeton, New Jersey

ETS Research Report No. RR-13-28

December 2013

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/

To obtain a copy of an ETS research report, please visit http://www.ets.org/research/contact.html

Action Editor: Shelby Haberman

Reviewers: Robert Mislevy and Wendy Yen

Copyright © 2013 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRE, LISTENING. LEARNING. LEADING., and TOEFL are registered trademarks of Educational Testing Service (ETS).

ADVANCED PLACEMENT is a registered trademark of the College Board.

Abstract

Few would doubt that ETS researchers have contributed more to the general topic of item response theory (IRT) than individuals from any other institution. In this report, we briefly review most of those contributions, dividing them into sections by decades of publication, beginning with early work by Fred Lord and Bert Green in the 1950s and ending with recent work that produced models involving complex structures and multiple dimensions.

Key words: item response theory

Foreword

Since its founding in 1947, ETS has conducted a significant and wide-ranging research program that has focused on, among other things, psychometric and statistical methodology; educational evaluation; performance assessment and scoring; large-scale assessment and evaluation; cognitive, developmental, personality, and social psychology; and education policy. This broad-based research program has helped build the science and practice of educational measurement, as well as inform policy debates.

In 2010, we began to synthesize these scientific and policy contributions, with the intention to release a series of reports sequentially over the course of the next few years. These reports constitute the *ETS R&D Scientific and Policy Contributions Series*.

In the eighth report in the series, James Carlson and Matthias von Davier look at the role that ETS researchers have played in developing item response theory (IRT), which is used almost universally in large-scale assessment programs around the world. IRT's popularity is largely due to the fact that an IRT model may be used to estimate parameters of test items and abilities of test takers, with the estimates of item difficulty parameters and test taker abilities placed on the same scale. ETS researchers have been at the forefront of contributing to IRT, starting with early work by Tucker and Lord in the 1940s and 1950s, which helped lay the groundwork for IRT. In the 1960s and 1970s such ETS researchers as Birnbaum, Ross, and Samejima contributed to the more complete development of IRT. In the 1980s, IRT work broadened greatly, with IRT methodology, computer programs, and linking and equating procedures developed further by Lord, Wingersky, Stocking, Pashley, Holland, and Mislevy among many others, at ETS. With the advent of the 1990s, IRT use expanded in operational testing programs, with ETS researchers such as Muraki, Carlson, Yamamoto, Yen, Chang, and Mazzeo contributing to advanced item response modeling. In the 21st century, ETS researchers such as Haberman, Sinharay, Xu, van Rijn, M. von Davier, and Rijmen have continued the long tradition of contributing to the development of IRT through explanatory and multidimensional IRT models. In short, the body of work by ETS staff has been instrumental in developing IRT, and ETS is committed to continuing to play a leading role in IRT research.

Carlson, a principal psychometrician at ETS, serves as a senior advisor to lead psychometricians at ETS on various projects. He is also the Daniel Eignor executive editor of the ETS Research Report series and oversees the peer review process for papers written by R&D staff. From 2008 through 2010, he was the editor of the *Journal of Educational Measurement,* and he has served as advisory editor for the *Journal of Educational Measurement* through the terms of three editors. He has also served on the editorial board of *Educational Measurement: Issues and Practice*. Previously, he was the assistant director for psychometrics for the National Assessment Governing Board. In that capacity, he was the board's chief technical expert on all matters related to the design of the methodology of NAEP assessments, with responsibility for advising the executive director and board members on the development and execution of guidelines and standards for the overall technical integrity of NAEP assessments. Before joining the NAGB staff, Carlson held various senior-level research scientist positions at ETS, ACT, and CTB/McGraw-Hill. These positions followed distinguished service as professor of education at the University of Pittsburgh and the University of Ottawa, where he taught graduate-level statistics and psychometrics, provided consultation for various groups and individuals, and supervised applied development, research, and evaluation projects.

As a research director at ETS, M. von Davier manages a group of researchers concerned with methodological questions arising in large-scale international comparative studies in education. He is the editor-in-chief of the *British Journal of Mathematical and Statistical Psychology* and one of the founding editors of the SpringerOpen journal *Large Scale Assessments in Education*, which is sponsored by the International Association for the Evaluation of Educational Achievement (IEA) and ETS through the IEA-ETS Research Institute (IERI). He is also an honorary senior research fellow at the University of Oxford. His current work at ETS involves the psychometric methodologies used in analyzing cognitive skills data and background data from large-scale educational surveys, such as the Organisation for Economic Co-operation and Development's PIAAC and PISA, as well as IEA's TIMSS and PIRLS. His work at ETS also involves the development of extensions and of estimation methods for multidimensional models for item response data and the improvement of models and estimation methods for the analysis of data from large-scale educational survey assessments. Prior to joining ETS, he led a research group on computer assisted science learning, was co-director of the "Computer as a tool for learning" section at the Institute for Science Education

(IPN) in Kiel, Germany, and was an associate member of the Psychometrics & Methodology Department of IPN. During his 10-year tenure at IPN, he developed commercially available software for analyses with the Rasch model, latent class analysis models, and mixture distribution Rasch models. He taught courses on foundations of neural networks and on psychometrics and educational psychology at the University of Kiel for the Department of Psychology as well as for the Department of Education.

Future reports in the *ETS R&D Scientific and Policy Contributions Series* will focus on other major areas of research and education policy in which ETS has played a role.

Ida Lawrence
Senior Vice-President
Research & Development Division
ETS

Table of Contents

Item response theory (IRT) models, in their many forms, are undoubtedly the most widely used models in large-scale operational assessment programs. They have grown from negligible usage prior to the 1980s to almost universal usage in large-scale assessment programs, not only in the United States, but in many other countries with active and up-to-date programs of research in the area of psychometrics and educational measurement.

Perhaps the most important feature leading to the dominance of IRT in operational programs is the characteristic of estimating individual item locations (difficulties) and test-taker locations (abilities) separately, but on the same scale, a feature not possible with classical measurement models. This estimation allows for tailoring tests through judicious item selection to achieve precise measurement for individual test takers (e.g., in computerized adaptive testing, CAT) or for important cut points on an assessment scale. It also provides mechanisms for placing different test forms on the same scale (linking and equating). Another important characteristic of IRT models is local independence: for a given location of test takers on the scale, the probablity of success on any item is independent of that of every other item on that scale. This characteristic is the basis of the lilkelihood function used to estimate test takers' locations on the scale.

Few would doubt that ETS researchers have contributed more to the general topic of IRT than individuals from any other institution. In this report we briefly review most of those contributions, dividing them into sections by decades of publication. Of course, many individuals in the field have changed positions between different testing agencies and universities over the years, some having been at ETS during more than one period of time. This report includes some contributions made by ETS researchers before taking a position at ETS, and some contributions made by researchers while at ETS, although they have since left. It is also important to note that IRT developments at ETS were not made in isolation. Many contributions were collaborations between ETS researchers and individuals from other institutions, as well as developments that arose from communications with others in the field.

Some Early Work Leading up to IRT (1940s and 1950s)

Ledyard Tucker[1] (1946) published a precursor to IRT in which he introduced the term *item characteristic curve*, using the normal ogive model (Green, 1980).[2] Green stated:

> Workers in IRT today are inclined to reference Birnbaum in Novick and Lord [sic] when needing historical perspective, but, of course Lord's 1955 monograph, done under Tuck's

> direction, precedes Birnbaum, and Tuck's 1946 paper precedes practically everybody. He used normal ogives for item characteristic curves, as Lord did later. (p. 4)

Some of the earliest work leading up to a complete specification of IRT was carried out at ETS during the 1950s by **Fred Lord** and **Bert Green**. Green was one of the first two psychometric fellows in the joint doctoral program of ETS and Princeton University. Note that the work of Lord and Green was completed prior to Rasch's (1960) publication describing and demonstrating the one-parameter IRT model, although in his preface Rasch mentions modeling data in the mid-1950s, leading to what is now referred to as the Rasch model. Further background on the statistical and psychometric underpinnings of IRT can be found in the work of a variety of authors, both at and outside of ETS (Bock, 1997; Green, 1980; Lord, 1952a, 1952b/1953, 1952c).[3]

Lord (1951, 1952a, 1952b/1953) discussed test theory that can be considered some of the earliest work in IRT. He used and explained many of the now common IRT terms such as item characteristic curves (ICCs), test characteristic curves (TCCs), and standard errors conditional on latent ability. [4] He also discussed what we now refer to as local independence and the invariance of item parameters (not dependent on the ability distribution of the test takers). His 1953 article is an excellent presentation of the basics of IRT, and he also mentions the relevance of works specifying mathematical forms of ICCs in the 1940s (by Lawley, by Mosier, and by Tucker), and in the 1950s, (by Carroll, by Cronbach & Warrington, and by Lazarsfeld).

The emphasis of Green (1950a/1951a, 1950b, 1950c, 1950d/1952, 1951b) was on analyzing item response data using latent structure (LS) and latent class (LC) models. Green (1951b) stated:

> Latent Structure Analysis is here defined as a mathematical model for describing the interrelationships of items in a psychological test or questionnaire on the basis of which it is possible to make some inferences about hypothetical fundamental variables assumed to underlie the responses. It is also possible to consider the distribution of respondents on these underlying variables. This study was undertaken to attempt to develop a general procedure for applying a specific variant of the latent structure model, the latent class model, to data. (abstract)

He also showed the relationship of the latent structure model to factor analysis (FA)

> The general model of latent structure analysis is presented, as well as several more specific models. The generalization of these models to continuous manifest data is indicated. It is noted that in one case, the generalization resulted in the fundamental equation of linear multiple factor analysis. (1950d, abstract)

The work of Green and Lord is significant for many reasons. An important one is that IRT (previously referred to as latent trait, or LT, theory) was shown by Green to be directly related to the models he developed and discussed. Lord (1952a) showed that if a single latent trait is normally distributed, fitting a linear FA model to the tetrachoric correlations of the items yields a unidimensional normal-ogive model for the item response function.

More Complete Development of IRT (1960s and 1970s)

During the 1960s and 1970s, Lord (1964a/1965b, 1964b, 1965a, 1967/1968a, 1968b/1970, 1968c) expanded on his earlier work to develop IRT more completely, and also demonstrated its use on operational test scores (including early software to estimate the parameters). Also at this time in two ETS Research Bulletins (RBs), **Allan Birnbaum** (1967) presented the theory of logistic models and **John Ross** (1965) studied how actual item response data fit Birnbaum's model. In another ETS RB, **Fumiko Samejima** (1968, 1969)[5] published her development of the graded response (GR) model suitable for polytomous data. The theoretical developments of the 1960s culminated in some of the most important work on IRT during this period, much of it assembled into Lord and Novick's (1968) *Statistical Theories of Mental Test Scores* (which also includes contributions of Birnbaum: Chapters 17, 18, 19, and 20). Also Samejima's continuing work on graded response models, begun in her research bulletin, was further developed (1972) while she held academic positions.

An important aspect of the work at ETS in the 1960s was the development of software, particularly by **Marilyn Wingersky,** Lord**,** and **Erling Andersen** (Andersen, 1972; Lord, 1967/1968a;[6] Lord & Wingersky, 1973) enabling practical applications of IRT. The LOGIST computer program (Lord, Wingersky, & Wood, 1976; see also Wingersky, 1983) was the standard IRT estimation software used for many years in many other institutions besides ETS. Lord (1975b) also published a report in which he evaluated LOGIST estimates using artificial data. Developments during the 1950s were limited by a lack of such software and computers sufficiently powerful to carry out the estimation of parameters. In his 1967 and 1968 publications, Lord presented a description and demonstration of the use of maximum likelihood (ML) estimation of the ability and item parameters in the three-parameter logistic (3PL) model, using SAT® items. He stated, with respect to ICCs:

> The problems of estimating such a curve for each of a large number of items simultaneously is one of the problems that has delayed practical application of Birnbaum's models since they were first developed in 1957. The first step in the present project (see Appendix B) was to devise methods for estimating three descriptive parameters simultaneously for each item in the Verbal test. (1968a, p. 992)

Lord also discussed and demonstrated many other psychometric concepts, many of which were not put into practice until fairly recently due to the lack of computing power and algorithms. In two publications (1965a, 1965b) he emphasized that ICCs are the functions relating probability of response to the underlying latent trait, not to the total test score, and that the former and not the latter can follow a cumulative normal or logistic function (a point he originally made much earlier, Lord, 1953). He also discussed (1967/1968a) optimum weighting in scoring and information functions of items from a Verbal SAT test form, as well as test information, and relative efficiency of tests composed of item sets having different psychometric properties. A very interesting fact is that Lord (1968a, p. 1004) introduced and illustrated multistage tests (MTs), and discussed their increased efficiency relative to "the present Verbal SAT" (p. 1005). What we now refer to as *router* tests in using MTs, Lord called *foretests*. He also introduced *tailor-made tests* in this publication (and in Lord, 1968c) and discussed how they would be administered using computers. Tailor-made tests are now, of course, commonly known as computerized adaptive tests (CATs); as suggested above, MTs and CATs were not employed in operational testing programs until fairly recently, but it is fascinating to note how long ago Lord introduced these notions and discussed and demonstrated the potential increase in efficiency of assessments achievable with their use. With respect to CATs Lord stated:

> The detailed strategy for selecting a sequence of items that will yield the most information about the ability of a given examinee has not yet been worked out. It should be possible to work out such a strategy on the basis of a mathematical model such as that used here, however. (1968a, p. 1005)

In this work, Lord also presented a very interesting discussion (1968a, p. 1007) on improving validity by using the methods described and illustrated. Finally, in the appendix, Lord derived the ML estimators (MLEs) of the item parameters and, interestingly points out the fact, well known today, that MLEs of the 3PL lower asymptote or *c* parameter, are often "poorly determined by the data" (p. 1014). As a result, he fixed these parameters for the easier items in carrying out his analyses.

During the 1970s Lord produced a phenomenal number of publications, many of them related to IRT, but many on other psychometric topics. On the topics related to IRT alone, he produced six publications besides those mentioned above; these publications dealt with such diverse topics as individualized testing (1972a/1974b), estimating power scores from tests that used improperly timed administration (1972b/1973b), estimating ability and item parameters with missing responses (1973a/1974a), the ability scale (1974c/1975c), practical applications of item characeristic curves (1977a/1977b), and equating methods (1975a). In perusing Lord's work, including Lord and Novick (1968), the reader should keep in mind that he discussed many item response methods and functions using classical test theory (CTT) as well as what we now call IRT. Other work by Lord includes discussions of item characteristic curves and information functions without, for example, using normal ogive or logistic IRT terminology, but the methodology he presented dealt with the theory of item response data. During this period, Andersen visited ETS and during his stay developed one of the seminal papers on testing goodness of fit for the Rasch model (Andersen, 1973). Besides the work of Lord, during this period ETS staff produced many publications dealing with IRT, both methodological and application oriented. **Gary Marco** (1977), for example, described three studies indicating how IRT can be used to solve three relatively intractable testing problems: designing a multipurpose test, evaluating a multistage test, and equating test forms using pretest statistics. He used data from various College Board testing programs and demonstated the use of the information function and relative efficiency using IRT for preequating. **Linda Cook** (Hambleton & Cook, 1977) coauthored an article on using LT models to analyze educational test data. Hambleton and Cook described a number of different IRT models and functions useful in practical applications, demonstrated their use, and cited computer programs that could be used in estimating the parameters. **Charles Kreitzberg**, **Martha Stocking**, and **Len Swanson** (1977) discussed potential advantages of CAT, constraints and operational requirements, psychometric and technical developments that make it practical, and its advantages over conventional paper-and-pencil testing. **Michael Waller** (1976) described a method of estimating Rasch model parameters eliminating the effects of random guessing, without using a computer, and reported a Monte Carlo study of performance of the method.

Broadening the Research and Application of IRT (the 1980s)

During this decade, psychometricians, with leadership from Fred Lord, continued to develop the IRT methodology. Also, of course, computer programs for IRT were further developed. During this time many ETS measurement professionals were engaged in assessing the use of IRT models for scaling dichotomous item response data in operational testing programs. In many programs, IRT linking and equating procedures were compared with conventional methods, to inform programs about whether changing these methods should be considered.

Further Developments and Evaluation of IRT Models

In this section we describe further psychometric developments at ETS, as well as research studies evaluating the models, using both actual test and simulated data.

Lord continued to contribute to IRT methodology with works by himself as well as coauthoring works dealing with unbiased estimators of ability parameters and their parallel forms reliability (1981b), a four-parameter logistic model (Barton & Lord, 1981), standard errors of IRT equating (1981a/1982b), IRT parameter estimation with missing data (1982a/1983a), sampling variances and covariances of IRT parameter estimates (Lord &Wingersky, 1982), IRT equating (Stocking & Lord, 1982/1983), statistical bias in ML estimation of IRT item parameters (1982c/1983c), estimating the Rasch model when sample sizes are small (1983b), comparison of equating methods (Lord & Wingersky, 1983b, 1984), reducing sampling error (Lord & Wingersky, 1983a; Wingersky & Lord, 1984), conjunctive and disjunctive item response functions (1984a), ML and Bayesian parameter estimation in IRT (1984b/1986), and confidence bands for item response curves with **Peter Pashley** (Lord & Pashley, 1988).

Although Lord was undoubtedly the most prolific ETS contributor to IRT during this period, other ETS staff members made many contributions to IRT. **Paul Holland** (1980), for example, wrote on the question, “When are IRT models consistent with observed data?” and Cressie and Holland (1981) examined how to characterize the manifest probabilities in LT models. Holland and **Paul Rosenbaum** (1985/1986) studied monotone unidimensional latent variable models. They discussed applications and generalizations and provided a numerical example. Holland (1987/1990b) also discussed the *Dutch identity* as a useful tool for studying IRT models and conjectured that a quadratic form based on the identity is a limiting form for log manifest probabilities for all smooth IRT models as test length tends to infinity (but see Zhang and Stout, 1997, later in this report). **Doug Jones** discussed the adequacy of LT models (1980) and robustness tools for IRT (1982).

Howard Wainer and several colleagues published articles dealing with standard errors in IRT (Wainer, 1981; Wainer & Thissen, 1982), review of estimation in the Rasch model for "longish tests" (Gustafsson, Morgan, & Wainer, 1980), fitting ICCs with spline functions (Thissen, Wainer & Winsberg, 1984), estimating ability with wrong models and inaccurate parameters (Jones, Kaplan, & Wainer, 1984), evaluating simulation results of IRT ability estimation (Rubin, Thissen & Wainer, 1984; Thissen & Wainer, 1984), and confidence envelopes for IRT (Thissen & Wainer, 1983). Wainer (1983) also published an article discussing IRT and CAT, which he described as a coming technological revolution. Thissen and Wainer (1985) followed up on Lord's earlier work, discussing the estimation of the *c* parameter in IRT. Wainer and Thissen (1987) used the 1PL, 2PL, and 3PL models to fit simulated data and study accuracy and efficiency of robust estimators of ability. For short tests, simple models and robust estimators best fit the data, and for longer tests more complex models fit well, but using robust estimation with Bayesian priors resulted in substantial shrinkage. Testlet theory was the subject of Wainer and Lewis (1989).

Bob Mislevy has also made numerous contributions to IRT, introducing Bayes modal estimation (1985a/1986b) in 1PL, 2PL, and 3PL IRT models, providing details of an expectation-maximization (EM) algorithm using two-stage modal priors, and in a simulation study, demonstrated improvement in estimation. Additionally he wrote on Bayesian treatment of latent variables in sample surveys (Mislevy 1985b, 1986a). Most significantly, Mislevy (1984) developed the first version of a model that would later become the standard analytic approach for the National Assessment of Educational Progress (NAEP) and virtually all other large scale international survey assessments (see also Beaton & Barone's [2013] history report and the report by Kirsch, Lennon, von Davier, & Yamamoto [in press] on the history of adult literacy assessments at ETS). Mislevy (1986c/1987a) also introduced application of empirical Bayes procedures, using auxililary information about test takers, to increase the precision of item parameter estimates. He illustrated the procedures with data from the Profile of American Youth survey. He also wrote (1987b/1988a) on using auxilliary information about items to estimate Rasch model item difficulty parameters and authored and coauthored other papers, several with **Kathy Sheehan**, dealing with use of auxiliary/collateral information with Bayesian procedures for estimation in IRT models (Mislevy, 1988b; Mislevy & Sheehan, 1988c/1989c). Another contribution Mislevy made (1986d) is a comprehensive discussion of FA models for test item data with reference to relationships to IRT models and work on extending currently available models. Mislevy and Sheehan (1988a, 1988b/1989a) discussed consequences of uncertainty in IRT linking and the information matrix in latent variable models. Mislevy and Wu (1988) studied the effects of missing responses and discussed the implications for ability and item parameter estimation relating to alternate test forms, targeted testing, adaptive testing, time limits, and omitted responses. Mislevy also coauthored a book chapter describing a hierarchical IRT model (Mislevy & Bock, 1989).

Many other ETS staff members made important contributions. Jones (1984a, 1984b) used asymptotic theory to compute approximations to standard errors of Bayesian and robust estimators studied by Wainer and Thissen. Rosenbaum wrote on testing the local independence assumption (1984b) and showed (1984a) that the observable distributions of item responses must satisfy certain constraints when two groups of examinees have generally different ability to respond correctly under a unidimensional IRT model. **Neil Dorans** (1985) contributed a book chapter on item parameter invariance. Douglass, Marco, and Wingersky (1985) studied the use of approximations to the 3PL model in item parameter estimation and equating. Methodology for comparing distributions of item responses for two groups was contributed by Rosenbaum (1985). **Rob McKinley** and **Craig Mills** (1985) compared goodness of fit statistics in IRT models, and **Neal Kingston** and Dorans (1985) explored item-ability regressions as a tool for model fit.

Kumi Tatsuoka (1986) used IRT in developing a probabilistic model for diagnosing and classifying cognitive errors. While she held a postdoctoral fellowship at ETS, **Lynne Steinberg** coathored (Thissen & Steinberg, 1986) a widely used and cited taxonomy of IRT models, which mentions, among other contributions, that the expressions they use suggest additional, as yet undeveloped, models. One explicitly suggested is basically the two-parameter partial credit (2PPC) model developed by **Wendy Yen** (see Yen & Fitzpatrick, 2006) and the equivalent generalized partial credit (GPC) model developed by **Eiji Muraki** (1992a/1992b), both some years after the Thissen-Steinberg article. Rosenbaum (1987) developed and applied three nonparametric methods for comparisons of the shapes of two item characteristic surfaces. Stocking (1988a) developed two methods of on-line calibration for CAT tests and compared them in a simulation using item parameters from an operational assessment. She also (1988b) conducted a study on calibration using different ability distributions, concluding that the best estimation for applications that are highly dependent on item parameters, such as CAT and test construction, resulted when the calibration sample contained widely dispersed abilities. McKinley (1988) studied six methods of combining item parameter estimates from different samples using real and simulated item response data. He stated, "results support the use of covariance matrix-weighted averaging and a procedure that involves sample-size-weighted averaging of estimated item characteristic curves at the center of the ability distribution." (abstract). McKinley also (1989a) developed and evaluated with simulated data a confirmatory multidimensional IRT (MIRT) model. **Kentaro Yamamoto** (1989) developed HYBRID, a model combining IRT and LC analysis, and used it to "present a structure of cognition by a particular response vector or set of them" (abstract). The software developed by Yamamoto was also used in a paper by Mislevy & Verhelst (1990) that presented an attempt to identify latent groups of test takers. **Valery Folk** (Folk & Green, 1989) coauthored a work on adaptive estimation when the unidimensionality assumption of IRT is violated.

IRT Software Development and Evaluation

With respect to IRT software, Mislevy and Stocking (1987) provided a guide to use of the LOGIST and BILOG computer programs that was very helpful to new users of IRT in applied settings. Mislevy, of course, was one of the developers of BILOG (Mislevy & Bock, 1983). Wingersky (1987), the primary developer of LOGIST, developed and evaluated, with real and artificial data, a one-stage version of LOGIST for use when estimates of item parameters but not test-taker abilities are required. Item parameter estimates were not as good as those from LOGIST, and the one-stage software did not reduce computer costs when there were missing data in the real dataset. Stocking (1989) conducted a study of estimation errors and relationship to properties of the test or item set being calibrated; she recommended improvements to the methods used in the LOGIST and BILOG programs. Yamamoto (1989) produced the Hybil software for the HYBRID model we referred to above. Both Hybil and BILOG utilize marginal ML estimation, whereas LOGIST uses joint ML estimation methods.

Explanation, Evaluation, and Application of IRT Models

During this decade ETS scientists began exploring the use of IRT models with operational test data and producing works explaining IRT models for potential users. Applications of IRT were seen in many ETS testing programs.

Lord's book, *Applications of Item Response Theory to Practical Testing Problems* (1980a), presented much of the current IRT theory in language easily understood by many practitioners. It covered basic concepts, comparison to CTT methods, relative efficiency, optimal number of choices per item, flexilevel tests, multistage tests, tailored testing, mastery testing, estimating ability and item parameters, equating, item bias, omitted responses, and estimating true score distributions. Lord (1980b) also contributed a book chapter on practical issues in tailored testing.

Isaac Bejar illustrated use of item characteristic curves in studying dimensionality (1980), and he and Wingersky (1981/1982) applied IRT to the Test of Standard Written English, concluding that using the 3PL model and IRT preequating "did not appear to present problems" (abstract). Kingston and Dorans (1982) applied IRT to the *GRE*® Aptitude Test, stating that "the most notable finding in the analytical equatings was the sensitivity of the precalibration design to practice effects on analytical items … this might present a problem for any equating design" (abstract). Dorans and Kingston (1982a) used IRT in the analysis of the effect of item position on test taker responding behavior. They also (1982b) compared IRT and conventional methods for equating the GRE Aptitude Test, assessing the reasonableness of the assumptions of item response theory for GRE item types and examinee populations, and finding that the IRT precalibration design was sensitive to practice effects on analytical items. In addition, Kingston and Dorans (1984) studied the effect of item location on IRT equating and adaptive testing, and Dorans and Kingston (1985) studied effects of violation of the unidimensionality assumption on estimation of ability and item parameters and on IRT equating with the GRE Verbal Test, concluding that there were two highly correlated verbal dimensions that had an effect on equating, but that the effect was slight. Kingston, **Linda Leary**, and **Larry Wightman** (1985) compared IRT to conventional equating of the Graduate Management Admission Test (GMAT) and concluded that violation of local independence of this test had little effect on the equating results (they cautioned that further study was necessary before using other IRT-based procedures with the test). Kingston and McKinley (1987) investigated using IRT equating for the GRE Subject Test in Mathematics and also studied the unidimensionality and model fit assumptions, concluding that the test was reasonably unidimensional and the 3PL model was a reasonable fit to the data.

Cook, **Dan Eignor, Nancy Petersen** and colleagues wrote several explanatory papers and conducted a number of studies of application of IRT on operational program data, studying assumptions of the models, and various aspects of estimation and equating (Cook, Dorans, & Eignor, 1988; Cook, Dorans, Eignor, & Petersen, 1985; Cook & Eignor, 1985, 1989; Cook, Eignor, & Petersen, 1985; Cook, Eignor, & Schmitt, 1988; Cook, Eignor, & Stocking, 1988; Eignor, 1985). Cook, Eignor, and Taft (1985, 1988) examined effects of curriculum (comparing results for students tested before completing the curriculum with students tested after completing it) on stability of CTT and IRT difficulty parameter estimates, effects on equating, and the dimensionality of the tests. Cook, Eignor and Wingersky (1987), using simulated data based on actual SAT item parameter estimates, studied the effect of anchor item characteristics on IRT true-score equating.

Jones and Kreitzberg (1980) presented results of a study of CAT using the Broad-Range Tailored Test and concluded,"computerized adaptive testing is ready to take the first steps out of the laboratory environment and find its place in the educational community" (abstract). **Janice Scheuneman** (1980) produced a book chapter on LT theory and item bias**. Marilyn Hicks** (1983) compared IRT equating with fixed versus estimated parameters and three "conventional" equating methods using *TOEFL*® test data, concluding that fixing the *b* parameters to pretest values (essentially this is what we now call preequating) is a "very acceptable option." She followed up (1984) with another study in which she examined controlling for native language and found this adjustment resulted in increased stability for one test section but a decrease in another section. Peterson, Cook, and Stocking (1983) studied several equating methods using SAT data and found that for reasonably parallel tests, linear equating methods perform adequately, but when tests differ somewhat in content and length, methods based on the three-parameter logistic IRT model lead to greater stability of equating results. In a review of research on IRT and conventional equating procedures, Cook and Petersen (1987) discussed how equating methods are affected by sampling error, sample characteristics, and anchor item characteristics, providing much useful information for IRT users.

Cook coauthored a book chapter (Hambleton & Cook, 1983) on robustness of IRT models, including effects of test length and sample size on precision of ability estimates. Several ETS staff members contributed chapters to that same edited book on applications of item response theory (Hambleton, 1983). Bejar (1983) contributed an introduction to IRT and its assumptions; Wingersky (1983) a chapter on the LOGIST computer program; Cook and Eignor (1983) on practical considerations for using IRT in equating. Tatsuoka coauthored on appropriateness indices (Harnisch & Tatsuoka, 1983); and Yen wrote on developing a standardized test with the 3PL model (1983); both Tatsuoka and Yen later joined ETS.

Lord and **Cheryl Wild** (1985) compared the contribution of the four verbal item types to measurement accuracy of the GRE General Test, finding that the reading comprehension item type measures something slightly different from what is measured by sentence completion, analogy, or antonym item types. Dorans (1986) used IRT to study the effects of item deletion on equating functions and the score distribution on the SAT, concluding that reequating should be done when an item is dropped. Kingston and Holland (1986) compared equating errors using IRT and several other equating methods, and several equating designs, for equating the GRE General Test, with varying results depending on the specific design and method. Eignor and Stocking (1986a, 1986b) conducted two studies to investigate whether calibration or linking methods might be reasons for poor equating results on the SAT. In the first study they used actual data, and in the second they used simulations, concluding that a combination of differences in true mean ability and multidimensionality were consistent with the real data. Eignor, **Marna Golub-Smith**, and Wingersky (1986) studied the potential of a new plotting procedures for assessing fit to the 3PL model using SAT and TOEFL data. Sheehan and Wingersky (1986) also wrote on fit to IRT models, using regressions of item scores onto observed (number correct) scores rather than the previously used method of regressing onto estimated ability.

Bejar (1986/1990), using IRT, studied an approach to psychometric modeling that explicitly incorporates information on the mental models test takers use in solving an item, and concluded that it is not only workable, but also necessary for future developments in psychometrics. Kingston (1986) used full information FA to estimate difficulty and discrimination parameters of a MIRT model for the GMAT, finding there to be dominant first dimensions for both the quantitative and verbal measures. Mislevy (1987c) discussed implications of IRT developments for teacher certification. Mislevy (1989) presented a case for a new test theory combining modern cognitive psychology with modern IRT. Mislevy and Sheehan (1989b; also Sheehan & Mislevy, 1990) wrote on the integration of cognitive theory and IRT and illustrated their ideas using the Survey of Young Adult Literacy data. These ideas seem to be the first appearance of a line of research that continues today. The complexity of these models, built to integrate cognitive theory and IRT, evolved dramatically in the 21st century due to rapid increase in computational capabilities of modern computers and developments in understanding problem solving. **Ida Lawrence** coauthored a paper (Lawrence & Dorans, 1988) addressing the sample invariance properties of four equating methods with two types of test-taker samples (matched on anchor test score distributions or taken from different administrations and differing in ability). Results for IRT, Levine, and equipercentile methods differed for the two types of samples, whereas the Tucker observed score method did not. **Grant Henning** (1989) discussed the appropriateness of the Rasch model for multiple-choice data, in response to an article that questioned such appropriateness. McKinley (1989b) wrote an explanatory article for potential users of IRT. McKinley and **Gary Schaeffer** (1989) studied an IRT equating method for the GRE designed to reduce the overlap on test forms. Bejar, **Henry Braun**, and **Sybil Carlson** (1989), in a paper on methods used for patient management items in medical licensure testing, outlined recent developments and introduced a procedure that integrates those developments with IRT. **Robert Boldt** (1989) used LC analysis to study the dimensionality of the TOEFL and assess whether different dimensions were necessary to fit models to diverse groups of test takers. His findings were that a single dimension LT model fits TOEFL data well but "suggests the use of a restrictive assumption of proportionality of item response curves" (p. 123).

In 1983, ETS assumed the primary contract for NAEP, and ETS psychometricians were involved in designing analysis procedures, including the use of an IRT-based latent regression model using ML estimation of population parameters from observed item responses without estimating ability parameters for test takers (e.g., Mislevy, 1984, 1988c/1991). Asymptotic standard errors and tests of fit, as well as approximate solutions of the integrals involved, were developed in Mislevy's 1984 article. With leadership from **Sam Messick** (Messick, 1985;

Messick et al., 1983), a large team of ETS staff developed a complex assessment design involving new analysis procedures for direct estimation of average achievement of groups of students. **Rebecca Zwick** (1986, 1987) studied whether the NAEP reading data met the unidimensionality assumption underlying the IRT scaling procedures. Mislevy (1988c) wrote on making inferences about latent variables from complex samples, using IRT proficiency estimates as an example and illustrating with NAEP reading data. The innovations introduced include the linking of multiple test forms using IRT, a task that would be virtually impossible without IRT-based methods, as well as the intregration of IRT with a regression-based population model that allows the prediction of an ability prior, given background data collected in student questionnaires along with the cogntive NAEP tests.

Advanced Item Response Modeling: The 1990s

During the 1990s, the use of IRT in operational testing programs expanded considerably. IRT methodology for dichotomous item response data was well developed and widely used by the end of the 1980s. In the early years of the 1990s, models for polytomous item response data were developed and began to be used in operational programs. Muraki (1990) developed and illustrated an IRT model for fitting a polytomous item response theory model to Likert-type data. Muraki (1992a/1992b) also developed the GPC model, which has since become one of the most widely used models for polytomous IRT data. Concomitantly, before joining ETS, Yen[7] developed the 2PPC model that is identical to the GPC, differing only in the parameterization incorporated into the model. Muraki (1993a/1993b) also produced an article detailing the IRT information functions for the GPC model. **Hua-Hua Chang** and **John Mazzeo** (1994) discussed item category response functions (ICRFs) and the item response functions (IRFs), which are weighted sums of the ICRFs, of the partial credit and graded response models. They showed that if two polytomously scored items have the same IRF, they must have the same number of categories that have the same ICRFs. They also discussed theoretical and practical implications. Akkermans and Muraki (1997) studied and described characteristics of the item information and discrimination functions for partial credit items.

In work remeniscent of the earlier work of Green and Lord, **Drew Gitomer** and Yamamoto (1991a/1991b) described HYBRID (Yamamoto, 1989), a model that incorporates both LT and LC components; these authors, however, defined the latent classes by a cognitive analysis of the understanding that individuals have for a domain. Yamamoto and Everson (1997) also published a book chapter on this topic. **Randy Bennett**, Sebrechts, and Yamamoto (1991) studied new “cognitively sensitive measurement models,” analyzing them with the HYBRID model and comparing results to other IRT methodology, using partial-credit data from the GRE General Test. Works by Tatsuoka (1990, 1991) also contributed to the literature relating IRT to cognitive models. The integration of IRT and a person-fit measure as a basis for rule space, as proposed by Tatsuoka, allowed in-depth examinations of items that require multiple skills. Sheehan (1997a/1997b) developed a tree-based method of proficiency scaling and diagnostic assessment and applied it to developing diagnostic feedback for the SAT I Verbal Reasoning Test. Mislevy and Wilson (1996) presented a version of Wilson’s Saltus model, an IRT model that incorporates developmental stages that may involve discontinuities. They also demonstrated its use with simulated data and an example of mixed number subtraction.

The volume, *Test Theory for a New Generation of Tests* (Frederiksen, Mislevy, & Bejar, 1993), presented several IRT-based models that anticipated a more fully integrated approach providing information about measurement qualities of items as well as about complex latent variables that align with cognitive theory. Examples of these advances are the chapters by Yamamoto and Gitomer (1993) and Mislevy (1993a).

Eric Bradlow (1996) discussed the fact that, for certain values of item parameters and ability, the information about ability for the 3PL model will be negative and has consequences for estimation—a phenomenon that does not occur with the 2PL. **Peter Pashley** (1991) proposed an alternative to Birnbaum’s 3PL model in which the asymptote parameter is a linear component within the logit of the function. **Jinming Zhang** and Stout (1997) showed that Holland’s (1987/1990b) conjecture that a quadratic form for log manifest probabilities is a limiting form for all smooth unidimensional IRT models does not always hold; these authors provided counterexamples and suggested that only under strong assumptions can this conjecture be true.

Holland (1990a) published an article on the sampling theory foundations of IRT models. Stocking (1990) discussed determining optimum sampling of examinees for IRT parameter estimation. Chang and Stout (1993) showed that, for dichotomous IRT models, under very general and nonrestrictive nonparametric assumptions, the posterior distribution of examinee ability given dichotomous responses is approximately normal for a long test. Chang (1996) followed up with an article extending this work to polytomous responses, defining a global information function, and he showed the relationship of the latter to other information functions.

Mislevy (1991) published on randomization-based inference about latent variables from complex samples. Mislevy (1993b/1993c) also presented formulas for use with Bayesian ability estimates. While at ETS as a postdoctoral fellow, **Jim Roberts** coauthored works on the use of unfolding[8] (Laughlin & Roberts, 1996; Roberts & Laughlin, 1996). A parametric IRT model for unfolding dichotomously or polytomously scored responses, called the graded unfolding model (GUM), was developed; a subsequent recovery simulation showed that reasonably accurate estimates could be obtained with minimal data demands (e.g., as few as 100 subjects and 15 to 20 six-category items). The applicability of the GUM to common attitude testing situations was illustrated with real data on student attitudes toward capital punishment. Roberts, Donoghue, and Laughlin (1998/2000) described the generalized GUM (GGUM), which introduced a parameter to the model, allowing for variation in discrimination across items; they demonstrated the use of the model with real data.

Wainer and colleagues wrote further on testlet response theory, contributing to issues of reliability of testlet-based tests (Sireci, Thissen, & Wainer, 1991a/1991b). These authors also developed, and illustrated using operational data, statistical methodology for detecting differential item functioning (DIF) in testlets (Wainer, Sireci, & Thissen, 1991a/1991b). Thissen and Wainer (1990) also detailed and illustrated how *confidence envelopes* could be formed for IRT models. Bradlow, Wainer, and **Xiaohui Wang** (1998/1999) developed a Bayesian IRT model for testlets and compared results with those from standard IRT models using a released SAT dataset. They showed that degree of precision bias was a function of testlet effects and the testlet design. Sheehan and **Charlie Lewis** (1990/1992) introduced, and demonstrated with actual program data, a procedure for determining the effect of testlet nonequivalence on the operating characteristics of a computerized mastery test based on testlets.

Lewis and Sheehan (1990) wrote on using Bayesian decision theory to design computerized mastery tests. Contributions to CAT were made in a book, *Computer Adaptive Testing: A Primer*, edited by Wainer, Dorans, Flaugher, et al. (1990) with chapters by ETS psychometricians: “Introduction and History” (Wainer, 1990), “Item Response Theory, Item Calibration and Proficiency Estimation” (Wainer & Mislevy, 1990); “Scaling and Equating” (Dorans, 1990); “Testing Algorithms” (Thissen &, Mislevy,1990); “Validity” (Steinberg, Thissen, & Wainer 1990); “Item Pools” (Flaugher, 1990); and “Future Challenges” (Wainer, Dorans, Green, Mislevy, Steinberg, &Thissen, 1990). Automated item selection (AIS) using IRT was the topic of two publications (Stocking, Swanson, & Pearlman, 1991a, 1991b). Mislevy and Chang (1998/2000) introduced a term to the expression for probability of response vectors to deal with item selection in CAT, and to correct apparent incorrect response pattern probabilities in the context of adaptive testing. **Russell Almond** and Mislevy (1999) studied graphical modeling methods for making inferences about multifaceted skills and models in an IRT CAT environment, and illustrated in the context of language testing.

In an issue of an early volume of *Applied Measurement in Education*, Eignor, Stocking, and Cook (1990) expanded on their previous studies (Cook, Eignor, & Stocking, 1988) comparing IRT equating with several non-IRT methods and with different sampling designs. In another article in that same issue, Schmitt, Cook, Dorans, and Eignor (1990) reported on the sensitivity of equating results to sampling designs; Lawrence and Dorans (1990) contributed with a study of the effect of matching samples in equating with an anchor test; and **Skip Livingston**, Dorans and **David Wright** (1990) also contributed on sampling and equating methodolgy to this issue. Cook, Eignor, and Stocking (1990) also produced an ETS research report in which IRT and non-IRT equating methods were compared.

Zwick (1990) published an article showing when IRT and Mantel-Haenszel definitions of DIF coincide. Also in the DIF area, Dorans and Holland (1992) produced a widely disseminated and used work on the Mantel-Haenszel (MH) and standardization methodologies, in which they also detailed the relationship of the MH to IRT models. Their methodology, of course, is the mainstay of DIF analyses today, at ETS and at other institutions. Muraki (1999) described a stepwise DIF procedure based on the multiple group PC model. He illustrated the use of the model using NAEP writing trend data and also discussed item parameter drift. Pashley (1992) presented a graphical procedure, based on IRT, to display the location and magnitude of DIF along the ability continuum.

MIRT models, although developed earlier, were further developed and illustrated with operational data during this decade; McKinley coauthored an article (Reckase & McKinley, 1991) describing the discrimination parameter for these models. Muraki and **Jim Carlson** (1995) developed a multidimensional graded response (MGR) IRT model for polytomously scored items, based on Samejima's normal ogive GR model. Relationships to the Reckase-McKinley and FA models were discussed, and an example using NAEP reading data was presented and discussed. Zhang and Stout (1999a, 1999b) described models for detecting dimensionality and related them to FA and MIRT.

Lewis coauthored publications (McLeod & Lewis, 1999; McLeod, Lewis, & Thissen, 1999/2003) with a discussion of person-fit measures as potential ways of detecting memorization of items in a CAT environment using IRT, and introduced a new method. None of the three methods showed much power to detect memorization. Possible methods of altering a test when the model becomes inappropriate for a test taker were discussed.

IRT Software Development and Evaluation

During this period, Muraki developed the PARSCALE computer program (Muraki & Bock, 1993) that has become one of the most widely used IRT programs for polytomous item response data. At ETS it has been incorporated into the GENASYS software used in many operational programs to this day. Muraki (1992c) also developed the RESGEN software, also widely used, for generating simulated polytomous and dichotomous item response data.

Many of the research projects in the literature reviewed here involved development of software for estimation of newly developed or extended models. Some examples involve Yamamoto’s (1989) HYBRID model, the MGR model (Muraki & Carlson, 1995) for which Muraki created the POLYFACT software, and the Saltus model (Mislevy & Wilson, 1996) for which an EM algorithm-based program was created.

Explanation, Evaluation, and Application of IRT Models

In this decade ETS researchers continued to provide explanations of IRT models for users, to conduct research evaluating the models, and to use them in testing programs in which they had not been previously used. The latter activity is not emphasized in this section as it was for sections on previous decades because of the sheer volume of such work and the fact that it generally involves simply applying IRT to testing programs, whereas in previous decades the research made more of a contribution, with recommendations for practice in general. Although such work in the 1990s contributed to improving the methodology used in specific programs, it provided little information that can be generalized to other programs. This section, therefore covers research that is more generalizable, although illustrations may have used specific program data.

Some of this research provided new information about IRT scaling. **John Donoghue** (1992), for example, described the common misconception that the partial credit and GPC IRT model item category functions are symmetric, helping explain characteristics of items in these models for users of them. He also (1993) studied the information provided by polytomously scored NAEP reading items and made comparisons to information provided by dichotomously scored items, demonstrating how other users can use such information for their own programs. Donoghue and **Steve Isham** (1996/1998) used simulated data to compare IRT and other methods of detecting item parameter drift. Zwick (1991), illustrating with NAEP reading data, presented a discussion of issues relating to two questions: "What can be learned about the effects of item order and context on invariance of item parameter estimates?" and "Are common-item equating methods appropriate when measuring trends in educational growth?" Camili, Yamamoto, and **Ming-Mei Wang** (1993) studied scale shrinkage in vertical equating, comparing IRT with equipercentile methods using real data from NAEP and another testing program. Using IRT methods, variance decreased from fall to spring testings, and also from lower- to upper-grade levels, whereas variances have been observed to increase across grade levels for equipercentile equating. They discussed possible reasons for scale shrinkage and proposed a more comprehensive, model-based approach to establishing vertical scales. Yamamoto (1995) estimated IRT parameters using TOEFL data and his extended hybrid model (1989), which uses a combination of IRT and LC models to characterize when test takers switch from ability-based to random responses. He studied effects of time limits on speededness, finding that this model estimated the parameters more accurately than the usual IRT model. Everson and Yamamoto (1995), using three different sets of actual test data, found that the hybrid model successfully determined the switch point in the three datasets. **Mei Liu** coauthored (Lane, Stone, Ankenmann, & Liu, 1995) an article in which mathematics performance-item data were used to study the assumptions of and stability over time of item parameter estimates using the GR model. Mislevy and Sheehan (1994) used a tree-based analysis to examine the relationship of three types of item attributes (constructed-response [CR] vs. multiple choice [MC], surface features, aspects of the solution process) to operating characteristics (using 3PL parameter estimates) of computer-based *PRAXIS*™ mathematics items. Mislevy and Wu (1996) built on their previous research (1988) on estimation of ability when there are missing data due to assessment design (alternate forms, adaptive testing, targeted testing), focusing on using Bayesian and direct likelihood methods to estimate ability parameters.

Wainer, X.-B. Wang, and Thissen (1994) examined, in an IRT framework, the comparability of scores on tests in which test takers choose which CR prompts to respond to, and illustrated using the College Board *Advanced Placement*® Test in Chemistry.

Zwick, **Dotty Thayer**, and Wingersky (1995) studied the effect on DIF statistics of fitting a Rasch model to data generated with a 3PL model. The results, attributed to degredation of matching resulting from Rasch model ability estimation, indicated less sensitive DIF detection.

In 1992, special issues of the *Journal of Educational Measurement* and the *Journal of Educational Statistics* were devoted to methodology used by ETS in NAEP, including the NAEP IRT methodology. **Al Beaton** and **Eugene Johnson** (1992), and Mislevy, E. Johnson, and Muraki (1992) detailed how IRT is used and combined with the plausible values methodology to estimate proficiencies for NAEP reports. Mislevy, Beaton, **Bruce Kaplan**, and Sheehan (1992) wrote on how population characteristics are estimated from sparse matrix samples of item responses. Yamamoto and Mazzeo (1992) described IRT scale linking in NAEP.

IRT Contributions in the 21st Century

Advances in the Development of Explanatory and Multidimensional IRT Models

Multidimensional models and dimensionality considerations continued to be a subject of research at ETS, with many more contributions than in the previous decades. Zhang (2004a) proved that, when simple structure obtains, estimation of unidimensional or MIRT models by joint ML yields identical results, but not when marginal ML is used. He also conducted simulations and found that, with small numbers of items, MIRT yielded more accurate item parameter estimates but the unidimensional approach prevailed with larger numbers of items, and that when simple structure does not hold, the correlations among dimensions are overestimated.

A genetic algorithm was used by Zhang (2005b) in the maximization step of an EM algorithm to estimate parameters of a MIRT model with complex, rather than simple, structure. Simulated data suggested that this algorithm is a promising approach to estimation for this model. Zhang (2004b/2007) also extended the theory of conditional covariances to the case of polytomous items, providing a theoretical foundation for study of dimensionality. Several estimators of conditional covariance were constructed, including the case of complex incomplete designs such as those used in NAEP. He demonstrated use of the methodology with NAEP reading assessment data, showing that the dimensional structure is consistent with the purposes of reading that define NAEP scales, but that the degree of multidimensionality is weak in those data.

Shelby Haberman, **Matthias von Davier**, and **Yi-Hsuan Lee** (2008) showed that MIRT models can be based on ability distributions that are multivariate normal or multivariate polytomous, and showed, using empirical data, that under simple structure the two cases yield comparable results in terms of model fit, parameter estimates, and computing time. They also discussed numerical methods for use with the two cases.

Frank Rijmen wrote two papers dealing with methodology relating to MIRT models, further showing the relationship between IRT and FA models. As discussed in the first section of this report, such relationships were shown for more simple models by Bert Green and Fred Lord in the 1950s. In the first (2009a) paper, Rijmen showed how an approach to full information ML estimation can be placed into a graphical model framework, allowing for derivation of efficient estimation schemes in a fully automatic fashion. This avoids tedious derivations, and he demonstrated the approach with the bifactor and a MIRT model with a second-order dimension. In the second paper, (2009b/2010) Rijmen studied three MIRT models for testlet-based tests, showing that the second-order MIRT model is formally equivalent to the testlet model, which is a bifactor model with factor loadings on the specific dimensions restricted to being proportional to the loadings on the general factor.

M. von Davier and Carstensen (2007) edited a book dealing with multivariate and mixture distribution Rasch models, including extensions and applications of the models. Contributors to this book included: Haberman (2007b) on the interaction model; M. von Davier and Yamamoto (2007) on mixture distributions and hybrid Rasch models; Mislevy and Huang (2007) on measurement models as narrative structures; and Boughton and Yamamoto (2007) on a hybrid model for test speededness.

Tamas Antal (2007) presented a coordinate-free approach to MIRT models, emphasizing understanding these models as extensions of the univariate models. Based on earlier work by Rijmen, Tuerlinckx, de Boeck, and Kuppens (2003), Rijmen, Jeon, M. von Davier, and Rabe-

Hesketh (2013) described how MIRT models can be embedded and understood as special cases of generalized linear and nonlinear mixed models.

Haberman and **Sandip Sinharay** (2010a/2010b) studied the use of MIRT models in computing subscores, proposing a new statistical approach to examining when MIRT model subscores have added value over total number correct scores and subscores based on CTT. The MIRT-based methods were applied to several operational datasets, and results showed that these methods produce slightly more accurate scores than CTT-based methods.

Rose, M. von Davier, and **Xueli Xu** (2010) studied IRT modeling of nonignorable missing item responses in the context of large-scale international assessments, comparing using CTT and simple IRT models, the usual two treatments (missing item responses as wrong, or as not administered), with two MIRT models. One model used indicator variables as a dimension to designate where missing responses occurred, and the other was a multigroup MIRT model with grouping based on a within-country stratification by the amount of missing data. Using both simulated and operational data, they demonstrated that a simple IRT model ignoring missing data performed relatively well when the amount of missing data was moderate, and the MIRT-based models only outperformed the simple models with larger amounts of missingness, but they yielded estimates of the correlation of missingness with ability estimates and improved the reliability of the latter.

Peter van Rijn and Rijmen (2012) provided an excellent explanation of a "paradox" that in some MIRT models answering an additional item correctly can result in a decrease in the test taker's score on one of the latent variables, previously discussed in the psychometric literature. These authors show clearly how it occurs and also point out that it does not occur in testlet (restricted bifactor) models.

ETS researchers also continued to develop CAT methodology. **Duanli Yan**, Lewis, and Stocking (2004) introduced a nonparametric tree-based algorithm for adaptive testing and showed that it may be superior to conventional IRT methods when the IRT assumptions are not met, particularly in the presence of multidimensionality. While at ETS, **Alex Weissman** coauthored an article (Belov, Armstrong, & Weissman, 2008) in which a new CAT algorithm was developed and tested in a simulation using operational test data. Belov et al. showed that their algorithm, compared to another algorithm incorporating content constraints had lower maximum item exposure rates, higher utilization of the item pool, and more robust ability estimates when high (low) ability test takers performed poorly (well) at the beginning of testing.

The second edition of *Computerized Adaptive Testing: A Primer* (Wainer, Dorans, Eignor et al., 2000) was published and, as in the first edition (Wainer, Dorans, Flaugher, et al., 1990), many chapters were authored or coauthored by ETS researchers (Dorans, 2000; Flaugher, 2000; Steinberg, Thissen & Wainer, 2000; Thissen & Mislevy, 2000; Wainer, 2000; Wainer, Dorans, Green, Mislevy, Steinberg, & Thissen, 2000; Wainer & Eignor, 2000; Wainer & Mislevy, 2000). Xu and Douglas (2006) explored the use of nonparametric IRT models in CAT; derivatives of ICCs required by the Fisher information criterion might not exist for these models, so alternatives based on Shannon entropy and Kullback-Leibler information (which do not require derivatives) were proposed. For long tests these methods are equivalent to the maximum Fisher information criterion, and simulations showed them to perform similarly, and much better than random selection of items.

Diagnostic models for assessment including cognitive diagnostic (CD) assessment, as well as providing diagnostic information from common IRT models, continued to be an area of research by ETS staff. Yan, Almond, and Mislevy (2004), using a mixed number subtraction dataset, and cognitive research originally developed by Tatsuoka and her colleagues, compared several models for providing diagnostic information on score reports, including IRT and other types of models, and characterized the kinds of problems for which each is suited. They provided a general Bayesian psychometric framework to provide a common language, making it easier to appreciate the differences. M. von Davier (2005/2008a) presented a class of general diagnostic (GD) models that can be estimated by marginal ML algorithms; that allow for both dichotomous and polytomous items, compensatory and noncompensatory models; and subsume many common models including univariate and multivariate Rasch models, 2PL, PC and GPC, Facets, and a variety of skill profile models. He demonstrated the model using simulated as well as TOEFL iBT data.

Xu (2007) studied monotonicity properties of the GD model and found that, like the GPC model, monotonicity obtains when slope parameters are restricted to be equal, but does not when this restriction is relaxed, although model fit is improved. She pointed out that trade offs between these two variants of the model should be considerred in practice. M. von Davier (2007a) extended the GD model to a hierarchical model and further extended it (2007b) to the mixture general diagnostic (MGD) model (see also M. von Davier, 2008b), which allows for estimation of diagnostic models in multiple known populations as well as discrete unknown, or not directly observed mixtures of populations.

Xu and M. von Davier (2006) used a MIRT model specified in the GD model framework with NAEP data and verified that the model could satisfactorily recover parameters from a sparse data matrix and could estimate group characteristics for large survey data. Results under both single and multiple group assumptions and comparison with the NAEP model results were also presented. The authors suggested that it is possible to conduct cognitive diagnosis for NAEP proficiency data. Xu and M. von Davier (2008b) extended the GD model, employing a log-linear model to reduce the number of parameters to be estimated in the latent skill distribution. They extended that model (2008a) to allow comparison of constrained versus nonconstrained parameters across multiple populations, illustrating with NAEP data.

M. von Davier, DiBello, and Yamamoto (2006/2008) discussed models for diagnosis that combine features of MIRT, FA, and LC models. Hartz and Roussos (2008)[9] wrote on the fusion model for skills diagnosis, indicating that the development of the model produced advancements in modeling, parameter estimation, model fitting methods, and model fit evaluation procedures. Simulation studies demonstrated the accuracy of the estimation procedure, and effectiveness of model fitting and model fit evaluation procedures. They concluded that the model is a promising tool for skills diagnosis that merits further research and development.

Linking and equating also continue to be important topics of ETS research. In this section the focus is research on IRT-based linking/equating methods. M. von Davier and **Alina von Davier** (2004/2007, 2010) presented a unified approach to IRT scale linking and transformation. Any linking procedure is viewed as a restriction on the item parameter space, and then rewriting the log-likelihood function together with implementation of a maximization procedure under linear or nonlinear restrictions accomplishes the linking. Xu and M. von Davier (2008c/2008d) developed an IRT linking approach for use with the GD model and applied the proposed approach to NAEP data. Holland and Hoskens (2002) developed an approach viewing CTT as a first-order version of IRT and the latter as detailed elaborations of CTT, deriving general results for the prediction of true scores from observed scores, leading to a new view of linking tests not designed to be linked. They illustrated the theory using simulated and actual test data. M. von Davier, Xu, and Carstensen (2011) presented a model that generalizes approaches by Andersen

(1985), and Embretson (1991), respectively, to utilize MIRT in a multiple-population longitudinal context to study individual and group-level learning trajectories.

Research on testlets continued to be a focus at ETS, as well as research involving item families. X. Wang, Bradlow, and Wainer (2002) extended the development of testlet models to tests comprising polytomously scored and/or dichotomously scored items, using a fully Bayesian method. They analyzed data from the Test of Spoken English (TSE) and the North Carolina Test of Computer Skills, concluding that the latter exhibited significant testlet effects, whereas the former did not. Sinharay, **Matt Johnson**, and **David Williamson** (2003) used a Bayesian hierarchical model to study item families, showing that the model can take into account the dependence structure built into the families, allowing for calibration of the family rather than the individual items. They introduced the family expected response function (FERF) to summarize the probability of a correct response to an item randomly generated from the family, and suggested a way to estimate the FERF.

Wainer and X. Wang (2000/2001) conducted a study in which TOEFL data were fitted to an IRT testlet model, and for comparative purposes to a 3PL model. They found that difficulty parameters were estimated well with either model, but discrimination and lower asymptote parameters were biased when conditional independence was incorrectly assumed. Wainer also coauthored book chapters explaining methodology for testlet models (Glas, Wainer, & Bradlow, 2000; Wainer, Bradlow, & Du, 2000).

Yanmei Li, **Shuhong Li**, and **Lin Wang** (2010) used both simulated data and operational program data to compare the parameter estimation, model fit, and estimated information of testlets comprising both dichotomous and polytomous items. The models compared were a standard 2PL/GPC model (ignoring local item dependence within testlets) and a general dichotomous/polytomous testlet model. Results of both the simulation and real data analyses showed little difference in parameter estimation but more difference in fit and information. For the operational data, they also made comparisons to a MIRT model under a simple structure constraint, and this model fit the data better than the other two models.

Roberts, Donoghue, and Laughlin (2002) in a continuation of their research on the GGUM, studied the characteristics of marginal ML and expected a posteriori (EAP) estimates of item and test-taker parameter estimates, respectively. They concluded from simulations that accurate estimates could be obtained for items using 750 to 1,000 test takers and for test takers using 15 to 20 items.

Checking assumptions, including the fit of IRT models to both the items and test takers of a test, is another area of research at ETS during this period. Sinharay and M. Johnson (2003) studied the fit of IRT models to dichotomous item response data in the framework of Bayesian posterior model checking. Using simulations, they studied a number of discrepancy measures and suggest graphical summaries as having a potential to become a useful psychometric tool. In further work on this model checking (Sinharay, 2003a, 2003b, 2005, 2006; Sinharay, M. Johnson, & Stern, 2006) they discussed the model-checking technique, and IRT model fit in general, extended some aspects of it, demonstrated it with simulations, and discussed practical applications. **Weiling Deng** coauthored (de la Torre & Deng, 2008) an article proposing a modification of the standardized log likelihood of the response vector measure of person fit in IRT models, taking into account test reliability and using resampling methods. Evaluating the method, they found type I error rates were close to the nominal and power was good, resulting in a conclusion that the method is a viable and promising approach.

Based on earlier work during a postdoctoral fellowship at ETS, M. von Davier and Molenaar (2003) presented a person-fit index for dichotomous and polytomous IRT and latent structure models. Sinharay and **Ying Lu** (2007/2008) studied the correlation between fit statistics and IRT parameter estimates; previous researchers had found such a correlation, which was a concern for practitioners. These authors studied some newer fit statistics not studied in the previous research, and found these new statistics not to be correlated with the item parameters. Haberman (2009b) discussed use of generalized residuals in the study of fit of 1PL and 2PL IRT models, illustrating with operational test data.

Mislevy and Sinharay coauthored an article (Levy, Mislevy, & Sinharay, 2009) on posterior predictive model checking, a flexible family of model-checking procedures, used as a tool for studying dimensionality in the context of IRT. Factors hypothesized to influence dimensionality and dimensionality assessment are couched in conditional covariance theory and conveyed via geometric representations of multidimensionality. Key findings of a simulation study included support for the hypothesized effects of the manipulated factors with regard to their influence on dimensionality assessment and the superiority of certain discrepancy measures for conducting posterior predictive model checking for dimensionality assessment.

Xu and **Yue Jia** (2011) studied the effects on item parameter estimation in Rasch and 2PL models of generating data from different ability distributions (normal distribution, several degrees of generalized skew normal distributions), and estimating parameters assuming these different distributions. Using simulations, they found for the Rasch model that the estimates were little affected by the fitting distribution, except for fitting a normal to an extremely skewed generating distribution; whereas for the 2PL this was true for distributions that were not extremely skewed, but there were computational problems (unspecified) that prevented study of extremely skewed distributions.

M. von Davier and Yamamoto (2003) extended the GPC model to enable its use with discrete mixture IRT models with partially missing mixture information. The model includes LC analysis and multigroup IRT models as special cases. An application to large-scale assessment mathematics data, with three school types as groups and 20% of the grouping data missing, was used to demonstrate the model.

M. von Davier and Sinharay (2009/2010) presented an application of a stochastic approximation EM algorithm using a Metropolis-Hastings sampler to estimate the parameters of an item response latent regression (LR) model. These models extend IRT to a two-level latent variable model in which covariates serve as predictors of the conditional distribution of ability. Applications to data from NAEP were presented, and results of the proposed method were compared to results obtained using the current operational procedures.

Haberman (2004) discussed joint and conditional ML estimation for the dichotomous Rasch model, explored conditions for consistency and asymptotic normality, investigated effects of model error, estimated errors of prediction, and developed generalized residuals. The same author (Haberman, 2005a) showed that if a parametric model for the ability distribution is not assumed, the 2PL and 3PL (but not 1PL) models have identifiability problems that impose restrictions on possible models for the ability distribution. Haberman (2005b) also showed that LC item response models with small numbers of classes are competitive with IRT models for the 1PL and 2PL cases, showing that computations are relatively simple under these conditions. In another report, Haberman (2006) applied adaptive quadrature to ML estimation for IRT models with normal ability distributions, indicating that this method may achieve significant gains in speed and accuracy over other methods.

Information about the ability variable when an IRT model has a latent class structure was the topic of Haberman (2007a) in another publication. He also discussed reliability estimates and sampling and provided examples. Expressions for bounds on log odds ratios involving pairs of items for unidimensional IRT models in general, and explicit bounds for 1PL and 2Pl models were derived by Haberman, Holland, and Sinharay (2007). The results were illustrated through an example of their use in a study of model-checking procedures. These bounds can provide an elementary basis for assessing goodness of fit of these models. In another publication, Haberman (2008) showed how reliability of an IRT scaled score can be estimated and that it may be obtained even though the IRT model may not be valid.

Zhang (2005a) used simulated data to investigate whether Lord's bias function and weighted likelihood estimation method for IRT ability with known item parameters would be effective in the case of unknown parameters, concluding that they may not be as effective in that case. He also presented algorithms and methods for obtaining the global maximum of a likelihood, or weighted likelihood (WL), function.

Lewis (2001) produced a chapter on expected response functions (ERFs) in which he discussed Bayesian methods for IRT estimation. Zhang and **Ting Lu** (2007) developed a new corrected weighted likelihood (CWL) function estimator of ability in IRT models based on the asymptotic formula of the WL estimator; they showed via simulation that the new estimator reduces bias in the ML and WL estimators, caused by failure to take into account uncertainty in item parameter estimates. Y.-H. Lee and Zhang (2008) further studied this estimator and Lewis' ERF estimator under various conditions of test length and amount of error in item parameter estimates. They found that the ERF reduced bias in ability estimation under all conditions and the CWL under certain conditions.

Sinharay coedited a volume on psychometrics in the *Handbook of Statistics* (Rao & Sinharay, 2007), and contributions included chapters by: M. von Davier, Sinharay, **Andreas Oranje**, and Beaton (2007) describing recent developments and future directions in NAEP statistical procedures; Haberman and M. von Davier (2007) on models for cognitively based skills; M. von Davier and Rost (2007) on mixture distribution IRT models; and M. Johnson, Sinharay and Bradlow (2007) on hierarchical IRT models.

Deping Li and Oranje (2007) compared a new method for approximating standard error of regression effects estimates within an IRT-based regression model, with the imputation-based estimator used in NAEP. The method is based on accounting for complex samples and finite populations by Taylor series linearization, and these authors formally defined a general method, and extended it to multiple dimensions. The new method was compared to the NAEP imputation-based method.

Antal and Oranje (2007) described an alternative numerical integration applicable to IRT and emphasized its potential use in estimation of the LR model of NAEP. D. Li, Oranje, and Jiang (2007) discussed parameter recovery and subpopulation proficiency estimation using the hierarchical latent regression (HLR) model and made comparisons with the LR model using simulations. They found the regression effect estimates were similar for the two models, but there were substantial differences in the residual variance estimates and standard errors, especially when there was large variation across clusters because a substantial portion of variance is unexplained in LR.

M. von Davier and Sinharay (2004) discussed stochastic estimation for the LR model, and Sinharay and M. von Davier (2005) extended a bivariate approach that represented the gold standard for estimation to allow estimation in more than two dimensions. M. von Davier and Sinharay (2007) presented a Robbins-Monro type stochastic approximation algorithm for LR IRT models and applied this approach to NAEP reading and mathematics data.

IRT Software Development and Evaluation

X. Wang, Bradlow, and Wainer (2001, 2005) produced SCORIGHT, a program for scoring tests composed of testlets. M. von Davier (2005) presented stand-alone software for multidimensional discrete latent trait (MDLT) models that is capable of marginal ML estimation for a variety of IRT, mixture IRT, and hierarchical IRT models, as well as the GD approach. Haberman (2005b) presented a stand-alone general software for MIRT models. Rijmen (2006) presented a MATLAB toolbox utilizing tools from graphical modeling and Bayesian networks that allows estimation of a range of MIRT models.

Explanation, Evaluation, and Application of IRT Models

For the fourth edition of *Educational Measurement* edited by Brennan, Yen and **Anne Fitzpatrick** (2006) contributed the chapter on IRT, providing a great deal of information useful to both practictioners and researchers. Although other ETS staff were authors or coauthors of chapters in this book, they did not focus on IRT methodology, per se.

Muraki, **Catherine Hombo (McClellan),** and **Yong-Won Lee** (2000) presented IRT methodology for psychometric procedures in the context of performance assessments, including description and comparison of many IRT and CTT procedures for scaling, linking, and equating. **Linda Tang** and Eignor (2001), in a simulation, studied whether CTT item statistics could be used as collateral information along with IRT calibration to reduce sample sizes for pretesting TOEFL items, and found that CTT statistics, as the only collateral information, would not do the job.

Don Rock and **Judy Pollack** (2002) investigated model-based methods (including IRT-based methods), and more traditional methods of measuring growth in prereading and reading at the kindergarten level, including comparisons between demographic groups. They concluded that the more traditional methods may yield uninformative if not incorrect results.

Scrams, Mislevy, and Sheehan (2002) studied use of item variants for continuous linear computer-based testing. Results showed that calibrated difficulty parameters of analogy and antonym items from the GRE General Test were very similar to those based on variant family information, and, using simulations, they showed that precision loss in ability estimation was less than 10% in using parameters estimated from expected response functions based only on variant family information.

A study comparing linear, fixed common item, and concurrent parameter estimation equating methods in capturing growth was conducted and reported by **Mike Jodoin**, Keller, and Swaminathan (2003). A. von Davier and Wilson (2005) studied the assumptions made at each step of calibration through IRT true-score equating and methods of checking whether the assumptions are met by a dataset. Operational data from the AP Calculus AB exam were used as an illustration. **Ourania Rotou**, **Liane Patsula**, **Manfred Steffen**, and **Saba Rizavi** (2007) compared the measurement precision, in terms of reliability and conditional standard error of measurement (CSEM), of multistage (MS), CAT, and linear tests, using 1PL, 2PL, and 3PL IRT models. They found the MS tests to be superior to CAT and linear tests for the 1PL and 2PL models, and performance of the MS and CAT to be about the same, but better than the linear for the 3PL case.

Yuming Liu, Schulz, and **Lei Yu** (2008) compared the bootstrap and Markov chain Monte Carlo (MCMC) methods of estimation in IRT true-score equating with simulations based on operational testing data. Patterns of standard error estimates for the two methods were similar, but the MCMC produced smaller bias and mean square errors of equating. **Guemin Lee** and Fitzpatrick (2008), using operational test data, compared IRT equating by the Stocking-Lord method with and without fixing the *c* parameters. Fixing the *c* parameters had little effect on parameter estimates of the nonanchor items, but a considerable effect at the lower end of the scale for the anchor items. They suggeted that practitioners consider using the fixed-*c* method.

A regression procedure was developed by Haberman (2009a) to simultaneously link a very large number of IRT parameter estimates obtained from a large number of test forms, where each form has been separately calibrated and where forms can be linked on a pairwise basis by means of common items. An application to 2PL and GPC model data was also presented. Xu, Douglas, and Lee (2010) presented two methods of using nonparametric IRT models in linking, illustrating with both simulated and operational datasets. In the simulation study, they showed that the proposed methods recover the true linking function when parametric models do not fit the data or when there is a large discrepancy in the populations.

Y. Li (2012), using simulated data, studied the effects, for a test with a small number of polytomous anchor items, of item parameter drift on TCC linking and IRT true-score equating. Results suggest that anchor length, number of items with drifting parameters, and magnitude of the drift affected the linking and equating results. The ability distributions of the groups had little effect on the linking and equating results. In general, excluding drifted polytomous anchor items resulted in an improvement in equating results.

D. Li, **Yanlin Jiang**, and A. von Davier (2012) conducted a simulation study of IRT equating of six forms of a test, comparing several equating transformation methods and separate versus concurrent item calibration. The characteristic curve methods yielded smaller biases and smaller sampling errors (or accumulation of errors over time) so the former were concluded to be superior to the latter and were recommended in practice.

Livingston (2006) described IRT methodology for item analysis in a book chapter in *Handbook of Test Development* (Downing & Haladyna, 2006). In the same publication, **Cathy Wendler** and **Michael Walker** (2006) discussed IRT methods of scoring, and **Tim Davey** and **Mary Pitoniak** (2006) discussed designing CATs, including use of IRT in scoring, calibration, and scaling.

Almond, DiBello, **Brad Moulder**, and **Diego Zapata-Rivera** (2007) described Bayesian network models and their application to IRT-based CD modeling. The paper, designed to encourage practitioners to learn to use these models, is aimed at a general educational measurement audience, does not use extensive technical detail, and presents examples.

The Signs of (IRT) Things to Come

The body of work that ETS staff has contributed to in the development and applications of IRT, MIRT, and comprehensive integrated models based on IRT has been documented in multiple published monographs and edited volumes. At the point of writing this report, the history is still in the making; there are three more edited volumes that would have not been possible without the contributions of ETS researchers reporting on the use of IRT in various applications. More specifically:

- *Handbook of Modern Item Response Theory, Volume 2* (edited by Wim van der Linden & Ronald Hambleton, published September 2013) contains chapters by Shelby Haberman, John Mazzeo, Robert J. Mislevy, **Tim Moses**, Frank Rijmen, Sandip Sinharay, and Matthias von Davier.
- *Computerized Multistage Testing: Theory and Applications* (edited by Duanli Yan, Alina von Davier, & Charlie Lewis, expected March 2014) will contain chapters by Isaac Bejar, **Brent Bridgeman**, **Henry Chen**, Shelby Haberman, **Sooyeon Kim**, **Ed Kulick**, Yi-Hsuan Lee, Charlie Lewis, **Longjuan Liang**, Skip Livingston, John Mazzeo, **Kevin Meara**, **Chris Mills**, Andreas Oranje, **Fred Robin**, Manfred Steffen, Peter van Rijn, Alina von Davier, Matthias von Davier, **Carolyn Wentzel**, Xueli Xu, Kentaro Yamamoto, Duanli Yan, and Rebecca Zwick.
- *Handbook of International Large Scale International Assessment* (edited by Leslie Rutkowski, Matthias von Davier, & David Rutkowski, published December 2013) contains chapters by Henry Chen, **Eugeneo Gonzalez**, John Mazzeo, Andreas Oranje, Frank Rijmen, Matthias von Davier, **Jonathan Weeks**, Kentaro Yamamoto, and **Lei Y**e.

Summary

Over the past six decades, ETS has pushed the envelope of modeling item response data using a variety of latent trait models that are commonly subsumed under the label IRT. Early developments, software tools, and applications allowed insight into the particular advantages of approaches that use item response functions to make inferences about individual differences on latent variables. ETS has not only provided theoretical developments, but has also shown, in large scale applications of IRT, how these methodologies can be used to perform scale linkages in complex assessment designs, and how to enhance reporting of results by providing a common scale and unbiased estimates of individual or group differences.

In the past two decades, IRT, with many contributions from ETS researchers, has become an even more useful tool. One main line of development has connected IRT to cognitive models and integrated measurement and structural modeling. This integration allows for studying questions that cannot be answered by secondary analyses using simple scores derived from IRT- or CTT-based approaches. More specifically, differential functioning of groups of items, the presence or absence of evidence that suggests that multiple diagnostic skill variables can be identified, and comparative assessment of different modeling approaches are part of what the most recent generation of multidimensional explanatory item response models can provide.

ETS will continue to provide cutting edge research and development on future IRT-based methodologies, and continues to play a leading role in the field, as documented by the fact that nine chapters of the *Handbook of Modern Item Response Theory, Volume 2* are authored by ETS staff. Also, of course, at any point in time, including the time of publication of this work, there are numerous research projects being conducted by ETS staff, and for which reports are being drafted, reviewed, or submitted for publication. By the time this work is published, there will undoubtedly be additional publications not included herein.

References[10]

Akkermans, W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62,* 569–578.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44,* 341–359.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223–237.

Andersen, E. B. (1972). *A computer program for solving a set of conditional maximum likelihood equations arising in the Rasch model for questionnaires* (Research Memorandum No. RM-72-06). Princeton, NJ: Educational Testing Service.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

• Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Antal, T. (2007). *On multidimensional item response theory: A coordinate-free approach* (Research Report No. RR-07-30). Princeton, NJ: Educational Testing Service.

Antal, T., & Oranje, A. (2007). *Adaptive numerical integration for item response theory* (Research Report No. RR-07-06). Princeton, NJ: Educational Testing Service.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Report No. RR-81-20). Princeton, NJ: Educational Testing Service.

Beaton, A. E., & Barone, J. (2013). *Large-scale assessment and evaluation at ETS* (R&D Scientific and Policy Contributions Series). Manuscript in preparation.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 29,* 163–175.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283–296.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 1–23). Vancouver, Canada: Educational Research Institute of British Columbia.

Bejar, I. I. (1986). *A psychometric analysis of a three-dimensional spatial task* (Research Report No. RR-86-19). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237–245.

Bejar, I. I., Braun, H. I., & Carlson, S. B. (1989). *Psychometric foundations of testing based on patient management problems* (Research Memorandum No. RM-89-02). Princeton, NJ: Educational Testing Service.

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

Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement, 6*, 309–325.

Belov, D., Armstrong, R. D., & Weissman, A. (2008). A Monte Carlo approach for adaptive testing with content constraints. *Applied Psychological Measurement, 32*, 431–446.

Bennett, R. E., Sebrechts, M. M., & Yamamoto, K. (1991). *Fitting new measurement models to GRE General Test constructed-response item data* (Research Report No. RR-91-60). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1967). *Statistical theory for logistic mental test models with a prior distribution of ability* (Research Bulletin No. RB-67-12). Princeton, NJ: Educational Testing Service.

• Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–33.

Boldt, R. F. (1989). Latent structure analysis of the Test of English as a Foreign Language. *Language Testing, 6*, 123–142.

Boughton, K., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 147–156). New York, NY: Springer Science & Business Media.

Bradlow, E. T. (1996). Negative information and the three-parameter logistic model. *Journal of Educational and Behavioral Statistics, 21*, 179–185.

Bradlow, E. T., Wainer, H., & Wang, X. (1998). *A Bayesian random effects model for testlets* (Research Report No. RR-98-03). Princeton, NJ: Educational Testing Service.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Camilli, G., Yamamoto, K., & Wang, M.-M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17,* 379–388.

Chang, H.-H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika, 61,* 445–463.

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391–404.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58,* 37–52.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics, 13,* 19–43.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (Research Report No. RR-85-30). Princeton, NJ: Educational Testing Service.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. Hambleton (Ed.), *Applications of item response theory* (pp. 175–195). Vancouver, Canada: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1985). *An investigation of the feasibility of applying item response theory to equate achievement tests* (Research Report No. RR-85-31). Princeton, NJ: Educational Testing Service.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13,* 161–173.

Cook, L. L., Eignor, D. R., & Petersen, N. S. (1985). *A study of the temporal stability of IRT item parameter estimates* (Research Report No. RR-85-45). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability* (Research Report No. RR-88-52). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Stocking, M. L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures* (Research Report No. RR-88-41). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Stocking, M. L. (1990). *The effects on observed-and true-score equating procedures of matching on a fallible criterion: A simulation with test variation* (Research Report No. RR-90-25). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1985). *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates* (Research Report No. RR-85-38). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25,* 31–45.

Cook, L. L., Eignor, D. R., & Wingersky, M. S. (1987*). Specifying the characteristics of linking items used for item response theory item calibration* (Research Bulletin No. RB-87-24). Princeton, NJ: Educational Testing Service.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225-244.

Cressie, N., & Holland, P. W. (1981). *Characterizing the manifest probabilities of latent trait models* (Research Report No. RR-81-54). Princeton, NJ: Educational Testing Service.

Davey, T., & Pitoniak, M. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 543–573). Mahwah, NJ: Erlbaum.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45,* 159–177.

Donoghue, J. R. (1992). *On a common misconception concerning the partial credit and generalized partial credit polytomous IRT models* (Research Memorandum No. RM-92-12). Princeton, NJ: Educational Testing Service.

Donoghue, J, R. (1993). *An empirical examination of the IRT information in polytomously scored reading items* (Research Report No. RR-93-12). Princeton, NJ: Educational Testing Service.

Donoghue, J. R., & Isham, S. P. (1996). *Comparing the effectiveness of procedures to detect item parameter drift* (Research Report No. RR-96-35). Princeton, NJ: Educational Testing Service.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33–51.

Dorans, N. J. (1985). Item parameter invariance: The cornerstone of item response theory. In K. M. Rowland, & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 3, pp. 55–78). Greenwich, CT: JAI Press.

Dorans, N. J. (1986). The impact of item deletion on equating conversions and reported score distributions. *Journal of Educational Measurement, 23,* 245–264.

Dorans, N. J. (1990). Scaling and equating. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 137–160). Hillsdale, NJ: Erlbaum.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 135–158). Mahwah, NJ: Erlbaum.

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (Research Report No. RR-92-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kingston, N. M. (1982a). *The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory* (Research Report No. RR-82-22). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kingston, N. M. (1982b). *The feasibility of using item response theory as a psychometric model for the GRE aptitude test* (Research Report No. RR-82-12). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE Verbal Scale. *Journal of Educational Measurement*, *22*, 249–262.

Douglass, J. B., Marco, G. L., & Wingersky, M. S. (1985). *An evaluation of three approximate item response theory models for equating test scores* (Research Report No. RR-85-46). Princeton, NJ: Educational Testing Service.

• Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahway, NJ: Erlbaum.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT Verbal and Mathematical sections* (Research Report No. RR-85-10). Princeton, NJ: Educational Testing Service.

Eignor, D. R., Golub-Smith, M. L., & Wingersky, M. S. (1986). *Application of a new goodness-of-fit plot procedure to SAT and TOEFL item type data* (Research Report No. RR-86-47). Princeton, NJ: Educational Testing Service.

Eignor, D. R., & Stocking, M. L. (1986a). *An investigation of possible causes for the inadequacy of IRT pre-equating* (Research Report No. RR-86-14). Princeton, NJ: Educational Testing Service.

Eignor, D. R., & Stocking, M. L. (1986b). *The impact of different ability distributions on IRT preequating* (Research Report No. RR-86-49). Princeton, NJ: Educational Testing Service.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education*, *3*, 37–52.

• Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Everson, H. T., & Yamamoto, K. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (Research Report No. RR-95-16). Princeton, NJ: Educational Testing Service.

Flaugher, R. (1990). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 41–63). Hillsdale, NJ: Erlbaum.

Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 37–59). Mahwah, NJ: Erlbaum.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373–389.

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

Gitomer, D. H., & Yamamoto, K. (1991a). *Performance modeling that integrates latent trait and class theory* (Research Report No. RR-91-01). Princeton, NJ: Educational Testing Service.

Gitomer, D. H., & Yamamoto, K. (1991b). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28,* 173–189.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Dordrecht, The Netherlands: Kluwer Academic Publishers.

• Green, B. F., Jr. (1950a). *A general solution for the latent class model of latent structure analysis* (Research Bulletin No. RB-50-38). Princeton, NJ: Educational Testing Service.

Green, B. F., Jr. (1950b). *A proposal for a comparative study of the measurement of attitude* (Research Memorandum No. RM-50-20). Princeton, NJ: Educational Testing Service.

Green, B. F., Jr. (1950c). *A proposal for an empirical evaluation of the latent class model of latent structure analysis* (Research Memorandum No. RM-50-26). Princeton, NJ: Educational Testing Service.

Green, B. F., Jr. (1950d). *Latent structure analysis and its relation to factor analysis* (Research Bulletin No. RB-50-65). Princeton, NJ: Educational Testing Service.

Green, B. F., Jr. (1951a). A general solution for the latent class model of latent structure analysis. *Psychometrika, 16,* 151–166.

Green, B. F., Jr. (1951b). *Latent class analysis: A general solution and an empirical evaluation* (Research Bulletin No. RB-51-15). Princeton, NJ: Educational Testing Service.

Green, B. F., Jr. (1952). Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association, 47,* 71–76.

• Green, B. R., Jr. (1980). *Ledyard R Tucker's affair with psychometrics: The first 45 years*. Paper presented at a special symposium in honor of Ledyard R Tucker, The University of Illinois, Champaign, IL.

Gustafsson, J.-E., Morgan, A. M. B., & Wainer, H. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests. *Journal of Educational Statistics, 5,* 35–64.

Haberman, S. J. (2004). *Joint and conditional maximum likelihood estimation for the Rasch model of binary responses* (Research Report No. RR-04-20). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2005a). *Identifiability of parameters in item response models with unconstrained ability distributions* (Research Report No. RR-05-24). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2005b). *Latent-class item response models* (Research Report No. RR-05-28). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (Research Report No. RR-06-29). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2007a). *The information a test provides on an ability parameter* (Research Report No. RR-07-18). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2007b). The interaction model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 201–216). New York, NY: Springer.

Haberman, S. J. (2008). *Reliability of scaled scores* (Research Report No. RR-08-70). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2009a). *Linking parameter estimates derived from an item response model through separate calibrations* (Research Report No. RR-09-40). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2009b). *Use of generalized residuals to examine goodness of fit of item response m*odels (Research Report No. RR-09-15). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Holland, P. W., & Sinharay, S. (2007). Limits on log odds ratios for unidimensional item response theory models. *Psychometrika, 72,* 551–561.

Haberman, S. J., & Sinharay, S. (2010a). *How can multivariate item response theory be used in reporting of subscores* (Research Report No. RR-10-09). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2010b). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209–227.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol 26. Psychometrics* (pp. 1031–1038). Amsterdam, The Netherlands: Elsevier.

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

• Hambleton, R. K. (1983). *Applications of item response theory*. Vancouver, Canada: Educational Research Institute of British Columbia.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75–96.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31–49). New York, NY: Academic Press.

• Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 104–122). Vancouver, Canada: Educational Research Institute of British Columbia.

Hartz, S., & Roussos, L. (2008). *The fusion model for skills diagnosis: Blending theory with practicality* (Research Report No. RR-08-71). Princeton, NJ: Educational Testing Service.

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26,* 91–97.

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. *Applied Psychological Measurement, 7,* 255–266.

Hicks, M. M. (1984). *A comparative study of methods of equating TOEFL test scores* (Research Report No. RR-84-20). Princeton, NJ: Educational Testing Service.

Holland, P. (1980). *When are item response theory models consistent with observed data?* (Program Statistics Research Technical Report No. 80-07). Princeton, NJ: Educational Testing Service.

Holland, P. (1987). *The Dutch identity: A new tool for the study of item response theory models* (Research Report No. RR-87-27). Princeton, NJ: Educational Testing Service.

Holland, P. (1990a). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Holland, P. (1990b). The Dutch identity: A new tool for the study of item response theory models. *Psychometrika, 55,* 5–18.

Holland, P. W., & Hoskens, M. (2002). *Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test* (Research Report No. RR-02-20). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Rosenbaum, P. R. (1985). *Conditional association and unidimensionality in monotone latent variable models* (Research Report No. RR-85-47). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14,* 1523–1543.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71,* 229–250.

Johnson, M., Sinharay, S., & Bradlow, E. T. (2007). Hierarchical item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 587–605). Amsterdam, The Netherlands, Elsevier.

Jones, D. H. (1980). *On the adequacy of latent trait models* (Program Statistics Research Technical Report No. 80-08). Princeton, NJ: Educational Testing Service.

Jones, D. H. (1982). *Tools of robustness for item response theory* (Research Report No. RR-82-41). Princeton, NJ: Educational Testing Service.

Jones, D. H. (1984a). *Asymptotic properties of the robustified jackknifed estimator* (Research Report No. RR-84-41). Princeton, NJ: Educational Testing Service.

Jones, D. H. (1984b). *Bayesian estimators, robust estimators: A comparison and some asymptotic results* (Research Report No. RR-84-42). Princeton, NJ: Educational Testing Service.

Jones, D. H., Kaplan, B. A., & Wainer, H. (1984). *Estimating ability with three item response models when the models are wrong and their parameters are inaccurate* (Research Report No. RR-84-26). Princeton, NJ: Educational Testing Service.

Jones, D. H., & Kreitzberg, C. B. (1980). *An empirical study of the Broad Range Tailored Test of Verbal Ability* (Research Report No. RR-80-05). Princeton, NJ: Educational Testing Service.

Kingston, N. M. (1986). *Assessing the dimensionality of the GMAT Verbal and Quantitative measures using full information factor analysis* (Research Report No. RR-86-13). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1982). *The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test* (Research Report No. RR-82-12). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147–154.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281–288.

Kingston, N. M., & Holland, P. W. (1986). *Alternative methods of equating the GRE General Test* (Research Report No. RR-86-16). Princeton, NJ: Educational Testing Service.

Kingston, N. M., Leary, L. F., & Wightman, L. E. (1985). *An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test* (Research Report No. RR-85-34). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & McKinley, R. L. (1987). *Exploring the use of IRT equating for the GRE Subject Test in Mathematics* (Research Report No. RR-87-21). Princeton, NJ: Educational Testing Service.

Kirsch, I., Lennon, M. L., von Davier, M., & Yamamoto, K. (in press). *Educational Testing Service's Large-scale assessments: A historical view* (R&D Scientific and Policy Contributions). Princeton, NJ: Educational Testing Service.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1977). *Computerized adaptive testing: The concepts and its potentials* (Research Memorandum No. RM-77-03). Princeton, NJ: Educational Testing Service.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8,* 313–340.

Laughlin, J. E., & Roberts, J. S. (1996). *The graded unfolding model: A unidimensional item response model for unfolding graded responses* (Research Report No. RR-96-16). Princeton, NJ: Educational Testing Service.

Lawrence, I. M., & Dorans, N. J. (1988). *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (Research Report No. RR-88-23). Princeton, NJ: Educational Testing Service.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3,* 19–36.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld , S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. 4. Measurement and prediction* (pp. 362–472). Princeton, NJ: Princeton University Press.

Lee, G., & Fitzpatrick, A. R. (2008). A new approach to test score equating using item response theory with fixed c-parameters. *Asia Pacific Education Review, 9,* 238–261.

Lee, Y.-H., & Zhang, J. (2008). *Comparing different approaches of bias correction for ability estimation in IRT models* (Research Report No. RR-08-13). Princeton, NJ: Educational Testing Service.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 519–537.

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 163–171). New York, NY: Springer-Verlag.

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

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement, 49,* 167–189.

Li, D., & Oranje, A. (2007). *Estimation of standard error of regression effects in latent regression models using Binder's linearization* (Research Report No. RR-07-09). Princeton, NJ: Educational Testing Service.

Li, D., Oranje, A., & Jiang, Y. (2007). *Parameter recovery and subpopulation proficiency estimation in hierarchical latent regression models* (Research Report No. RR-07-27). Princeton, NJ: Educational Testing Service.

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (Research Report No. RR-12-09). Princeton, NJ: Educational Testing Service.

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (Research Report No. RR-10-21). Princeton, NJ: Educational Testing Service.

Liu, Y., Schulz, E. M., & Yu, L. (2008). Standard error estimation of 3PL IRT true score equating with an MCMC method. *Journal of Educational and Behavioral Statistics, 33,* 257–278.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). Mahwah, NJ: Erlbaum.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73–95.

Lord, F. M. (1951). *A theory of test scores and their relation to the trait measured* (Research Bulletin No. RB-51-13). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1952a). *A theory of test scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation.

Lord, F. M. (1952b). *The relation of test score to the ability underlying the test* (Research Bulletin No. RB-52-10). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1952c). *The scale proposed for the academic ability test* (Research Memorandum No. RM-52-03). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–549.

Lord, F. M. (1964a). *A note on the normal ogive or logistic curve* (Research Bulletin No. RB-64-56). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1964b). *A strong true score theory, with applications* (Research Bulletin No. RB-64-19). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1965a). An empirical study of item-test regression. *Psychometrika, 30,* 373–376.

Lord, F. M. (1965b). A note on the normal ogive or logistic curve in item analysis. *Psychometrika, 30,* 371–372.

Lord, F. M. (1967). *An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model* (Research Bulletin No. RB-67-34). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1968a). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020.

Lord, F. M. (1968b). *Estimating item characteristic curves without knowledge of their mathematical form* (Research Bulletin No. RB-68-08). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1968c). *Some test theory for tailored testing* (Research Bulletin No. RB-68-38). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1970). Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika, 35,* 43–50.

Lord, F. M. (1972a). *Individualized testing and item characteristic curve theory* (Research Bulletin No. RB-72-50). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1972b). *Power scores estimated by item characteristic curves* (Research Bulletin No. RB-72-46). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1973a). *Estimation of latent ability and item parameters when there are omitted responses* (Research Bulletin No. RB-73-37). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1973b). Power scores estimated by item characteristic curves. *Educational and Psychological Measurement, 33,* 219–224.

Lord, F. M. (1974a). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

Lord, F. M. (1974b). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. II, pp. 106–126). San Francisco, CA: Freeman.

Lord, F. M. (1974c). *The ‘ability’ scale in item characteristic curve theory* (Research Bulletin No. RB-74-19). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1975a). *A survey of equating methods based on item characteristic curve theory* (Research Bulletin No. RB-75-13). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1975b). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin No. RB-75-33). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1975c). The ‘ability’ scale in item characteristic curve theory. *Psychometrika, 40,* 205–217.

Lord, F. M. (1977a). *Practical applications of item characteristic curve theory* (Research Bulletin No. RB-77-03). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1977b). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138.

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

Lord, F. M. (1980b). Some how and which for practical tailored testing. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter, (Eds.), *Psychometrics for educational debates* (pp. 189–205). New York, NY: Wiley.

Lord, F. M. (1981a). *Standard error of an equating by item response theory* (Research Report No. RR-81-49). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1981b). *Unbiased estimators of ability parameters, and of their parallel-forms reliability* (Research Report No. RR-81-50). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1982a). *Maximum likelihood estimation of item response parameters when some responses are omitted* (Research Report No. RR-82-05). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1982b). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6,* 463–472.

Lord, F. M. (1982c). *Statistical bias in maximum likelihood estimation of item parameters* (Research Report No. RR-82-20). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1983a). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477–482.

Lord, F. M. (1983b). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51–61). New York, NY: Academic Press.

Lord, F. M. (1983c). Statistical bias in maximum likelihood estimation of item parameters. *Psychometrika, 48*, 425–435.

Lord, F. M. (1984a). *Conjunctive and disjunctive item response functions* (Research Report No. RR-84-45). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1984b). *Maximum likelihood and Bayesian parameter estimation in item response theory* (Research Report No. RR-84-30). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

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

Lord, F. M., & Pashley, P. (1988). *Confidence bands for the three-parameter logistic item response curve* (Research Report No. RR-88-67). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Wild, C. L. (1985). *Contribution of verbal item types in the GRE General Test to accuracy of measurement of the verbal scores* (Research Report No. RR-85-29). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Wingersky, M. S. (1973). *A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. RM-73-02). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Wingersky, M. S. (1982). *Sampling variances and covariances of parameter estimates in item response theory* (Research Report No. RR-82-33). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Wingersky, M. *S*. (1983a). *An investigation of methods for reducing sampling error in certain IRT procedures* (Research Report No. RR-83-28). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Wingersky, M. S. (1983b). *Comparison of IRT observed-score and true-score 'equatings'* (Research Report No. RR-83-26). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8,* 453–461.

Lord, F. M., Wingersky, M. S., & Wood, R. L. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. RM-76-06). Princeton, NJ: Educational Testing Service.

Marco, G. L. (1977). *Item characteristic curve solutions to three intractable testing problems* (Research Bulletin No. RB-77-11). Princeton, NJ: Educational Testing Service.

McKinley, R. L. (1988). A comparison of six methods for combining multiple IRT item parameter estimates. *Journal of Educational Measurement, 25,* 233–246.

McKinley, R. L. (1989a). *Confirmatory analysis of test structure using multidimensional item response theory* (Research Report No. RR-89-31). Princeton, NJ: Educational Testing Service.

McKinley, R. L. (1989b). Methods plainly speaking: An introduction to item response theory. *Measurement and Evaluation in Counseling and Development, 22,* 37–57.

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

McKinley, R. L., & Schaeffer, G. A. (1989). *Reducing test form overlap of the GRE Subject Test in Mathematics using IRT triple-part equating* (Research Report No. RR-89-08). Princeton, NJ: Educational Testing Service.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23,* 147–160.

McLeod, L. D., Lewis, C., & Thissen, D. (1999). *A Bayesian method for the detection of item preknowledge in computerized adaptive testing* (LSAC Computerized Testing Report No. 98-07). Princeton, NJ: Educational Testing Service.

McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27,* 121–137.

Messick, S. J. (1985). *The 1986 NAEP design: Changes and challenges* (Research Memorandum No. RM-85-02). Princeton, NJ: Educational Testing Service.

Messick, S. J., Beaton, A. E., Lord, F. M., Baratz, J. C., Bennett, R. E., Duran, R. P., … Wainer, H. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report No. 83-01). Princeton, NJ: Educational Testing Service.

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

Mislevy, R. (1985a). *Bayes modal estimation in item response models* (Research Report No. RR-85-33). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1985b). *Inferences about latent populations from complex samples* (Research Report No. RR-85-41). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1986a). *A Bayesian treatment of latent variables in sample surveys* (Research Report No. RR-86-01). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1986b). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. (1986c). *Exploiting auxiliary information about examinees in the estimation of item parameters* (Research Report No. RR-86-18). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1986d). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3–31.

Mislevy, R. (1987a). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81–91.

Mislevy, R. (1987b). *Exploiting auxiliary information about items in the estimation of Rasch item parameters* (Research Report No. RR-87-26). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1987c). Recent developments in item response theory with implications for teacher certification. *Review of Research in Education, 14,* 239–275.

Mislevy, R. (1988a). Exploiting auxiliary information about items in the estimation of Rasch item parameters. *Applied Psychological Measurement, 12,* 281–296.

Mislevy, R. (1988b). *Exploiting collateral information in the estimation of item parameters* (Research Report No. RR-88-53). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1988c). *Randomization-based inferences about latent variables from complex samples* (Research Report No. RR-88-54). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1989). *Foundations of a new test theory* (Research Report No. RR-89-52). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R. (1993a). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Erlbaum.

Mislevy, R. (1993b). *Some formulas for use with Bayesian ability estimates* (Research Report No. RR-93-03). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1993c). Some formulas for use with Bayesian ability estimates. *Educational and Psychological Measurement, 53,* 315–328.

Mislevy, R. J., Beaton, A. E., Kaplan, B. A., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29,* 133–161.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.

Mislevy, R., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 57–75). San Diego, CA: Academic Press.

Mislevy, R., & Chang, H.-H. (1998). *Does adaptive testing violate local independence?* (Research Report No. RR-98-33). Princeton, NJ: Educational Testing Service.

Mislevy, R., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65*, 149–156.

Mislevy, R. J., & Huang, C.-W. (2007). Measurement models as narrative structures. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch model: Extensions and applications* (pp.15–35). New York, NY: Springer Science + Business Media, LLC, 2007

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131–154.

Mislevy, R. J., & Sheehan, K. M. (1988a). *Some consequences of the uncertainty in IRT linking procedures* (Research Report No. RR-88-38). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Sheehan, K. M. (1988b). *The information matrix in latent-variable models* (Research Report No. RR-88-24). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Sheehan, K. M. (1988c). *The role of collateral information about examinees in item parameter estimation* (Research Report No. RR-88-55). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Sheehan, K. M. (1989a). Information matrices in latent-variable models. *Journal of Educational Statistics, 14,* 335–350.

Mislevy, R. J., & Sheehan, K. M. (1989b). *Integrating cognitive and psychometric models to measure document literacy* (Research Report No. RR-89-51). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Sheehan, K. M. (1989c). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661–679.

Mislevy, R. J., & Sheehan, K. M. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills* (Research Report No. RR-94-14). Princeton, NJ: Educational Testing Service.

Mislevy, R., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG* (Research Report No. RR-87-43). Princeton, NJ: Educational Testing Service.

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

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71.

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report No. RR-88-48). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30). Princeton, NJ: Educational Testing Service.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14,* 59–71.

Muraki, E. (1992a). *A generalized partial credit model: Application of an EM algorithm* (Research Report No. RR-92-06). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992b). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16,* 159–176.

Muraki, E. (1992c). *RESGEN item response generator* (Research Report No. RR-92-07). Princeton, NJ: Educational Testing Service.

Muraki, E. (1993a). *Information functions of the generalized partial credit model* (Research Report No. RR-93-27). Princeton, NJ: Educational Testing Service.

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

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36*, 217–232.

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT-based test scoring and item analysis for graded items and rating scales* [Computer program]. Chicago, IL: Scientific Software.

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

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*, 325–337.

Pashley, P. J. (1991). *An alternative three-parameter logistic item response model* (Research Report No. RR-91-10). Princeton, NJ: Educational Testing Service.

Pashley, P. J. (1992). *Graphical IRT-based DIF analyses* (Research Report No. RR-92-66). Princeton, NJ: Educational Testing Service.

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Rao, C. R., & Sinharay, S. (Eds.). (2007). *Handbook of statistics: Vol. 26*. *Psychometrics.* Amsterdam, The Netherlands: Elsevier.

• Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, The Netherlands: Nielsen & Lydiche.

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

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* (Technical Report). Amsterdam, The Netherlands: VU University Medical Center.

Rijmen, F. (2009a). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Research Report No. RR-09-03). Princeton, NJ: Educational Testing Service.

Rijmen, F. (2009b). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison* (Research Report No. RR-09-37). Princeton, NJ: Educational Testing Service.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361–372.

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2013). A general psychometric approach for educational survey assessments: Flexible statistical models and efficient estimation methods. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis.* London, England: Chapman & Hall.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

• Roberts, J. S. (n.d.). *Item response theory models for unfolding*. Retrieved from http://www.psychology.gatech.edu/unfolding/Intro.html

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E (1998). *The generalized graded unfolding model: A general parametric item response model for unfolding graded responses* (Research Report No. RR-98-32). Princeton, NJ: Educational Testing Service.

Roberts, J. S., Donoghue, J. R., & Laughlin, L. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32.

Roberts, J. S., Donoghue, J. R., & Laughlin, L. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26,* 192–207.

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

Rock, D. A., & Pollack, J. M. (2002). *A model-based approach to measuring cognitive growth in pre-reading and reading skills during the kindergarten year* (Research Report No. RR-02-18). Princeton, NJ: Educational Testing Service.

Rose, N., von Davier, M., & Xu, X. (2010*). Modeling nonignorable missing data with item response theory (IRT)* (Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R. (1984a). *Are the item response patterns of two groups of examinees consistent with a difference in the distribution of a unidimensional latent variable?* (Research Report No. RR-84-27). Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R. (1984b). *Testing the local independence assumption in item response theory* (Research Report No. RR-84-09). Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R. (1985). Comparing distributions of item responses for two groups. *British Journal of Mathematical and Statistical Psychology, 38,* 206–215.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.

Ross, J. (1965). An *empirical study of a logistic mental test model* (Research Bulletin No. RB-65-11). Princeton, NJ: Educational Testing Service.

Rotou, O., Patsula, L. N., Steffen, M., & Rizavi, S. M. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests* (Research Report No. RR-07-04). Princeton, NJ: Educational Testing Service.

Rubin, D., Thissen, D., & Wainer, H. (1984). *A computer program for simulation evaluation of IRT ability estimators* (Research Report No. RR-84-37). Princeton, NJ: Educational Testing Service.

Samejima, F. (1968). *Estimation of latent ability using a response pattern of graded scores* (Research Bulletin No. RB-68-02). Princeton, NJ: Educational Testing Service.

• Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores *Psychometrika 34*(4, Whole Pt. 2).

• Samejima, F. (1972). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 37*(1, Whole Pt. 2).

Scheuneman, J. D. (1980). Latent trait theory and item bias. In L. J. Th. van der Kamp, W.F. Langerak, & D. N. M. de Gruijter, (Eds.), *Psychometrics for educational debates* (pp. 140–151). New York, NY: Wiley.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3,* 53–71.

Scrams, D. J., Mislevy, R. J., & Sheehan, K. M. (2002). *An analysis of similarities in item functioning within antonym and analogy variant families* (Research Report No. RR-02-13). Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (1997a). *A tree-based approach to proficiency scaling and diagnostic assessment* (Research Report No. RR-97-09). Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (1997b). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–352.

Sheehan, K. M., & Lewis, C. (1990). *Computerized mastery testing with nonequivalent testlets* (Research Report No. RR-90-16). Princeton, NJ: Educational Testing Service.

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

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

Sheehan, K. M., & Wingersky, M. S. (1986). *Using estimated item observed-score regressions to test goodness-of-fit of IRT models* (Research Report No. RR-86-23). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2003a). *Bayesian item fit analysis for dichotomous item response theory models* (Research Report No. RR-03-34). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2003b). *Practical applications of posterior predictive model checking for assessing fit of common item response theory models* (Research Report No. RR-03-33). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42,* 375–394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59,* 429–449.

Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (Research Report No. RR-03-28). Princeton, NJ: Educational Testing Service.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298–321.

Sinharay, S., Johnson, M. S., & Williamson, D. (2003). *An application of a Bayesian hierarchical model for item family calibration* (Research Report No. RR-03-04). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Lu, Y. (2007). *A further look at the correlation between item parameters and item fit statistics* (Research Report No. RR-07-36). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement, 45,* 1–15.

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BRGOUP program to higher dimensions* (Research Report No. RR-05-27). Princeton, NJ: Educational Testing Service.

Sireci, S. G., Thissen, D., & Wainer, H. (1991a). *On the reliability of testlet-based tests* (Research Report No. RR-91-22). Princeton, NJ: Educational Testing Service.

Sireci, S. G., Thissen, D., & Wainer, H. (1991b). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Steinberg, L., Thissen, D., & Wainer, H. (1990) Validity. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 187–231). Hillsdale, NJ: Erlbaum.

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 185–229). Mahwah, NJ: Erlbaum.

Stocking, M. L. (1988a). *Scale drift in on-line calibration* (Research Report No. RR-88-28). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1988b). *Specifying optimum examinees for item parameter estimation in item response theory* (Research Report No. RR-88-57). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1989). *Empirical estimation errors in item response theory as a function of test properties* (Research Report No. RR-89-05). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*, 461–475.

Stocking, M. L., & Lord, F. M. (1982). *Developing a common metric in item response theory* (Research Report No. RR-82-25). Princeton, NJ: Educational Testing Service.

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., & Pearlman, M. (1991a). *Automatic item selection (AIS) methods in the ETS testing environment* (Research Memorandum No. RM-91-05). Princeton, NJ: Educational Testing Service.

Stocking, M. L., Swanson, L., & Pearlman, M. (1991b). *Automated item selection using item response theory* (Research Report No. RR-91-09). Princeton, NJ: Educational Testing Service.

Tang, K. L., & Eignor, D. R. (2001). *A study of the use of collateral statistical information in attempting to reduce TOEFL IRT item parameter estimation sample sizes* (Research Report No. RR-01-11). Princeton, NJ: Educational Testing Service.

• Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. *Behaviormetrika, 13,* 73–85.

• Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1991). *A theory of IRT-based diagnostic testing* (Office of Naval Research Report). Princeton, NJ: Educational Testing Service.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 103–135). Hillsdale, NJ: Erlbaum.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–131). Mahwah, NJ: Erlbaum.

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

Thissen, D., & Wainer, H. (1983). *Confidence envelopes for item response theory* (Research Report No. RR-83-25). Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1984). *The graphical display of simulation results with applications to the comparison of robust IRT estimators of ability* (Research Report No. RR-84-36). Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1985). *Some supporting evidence for Lord's guideline for estimating "c" theory* (Research Report No. RR-85-15). Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics, 15,* 113–128.

Thissen, D., Wainer, H., & Winsberg, S. (1984). *Fitting item characteristic curves with spline functions* (Research Report No. RR-84-40). Princeton, NJ: Educational Testing Service.

• Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1–13.

Van Rijn, P., & Rijmen, F. (2012). *A note on explaining away and paradoxical results in multidimensional item response theory* (Research Report No. RR-12-13). Princeton, NJ: Educational Testing Service.

von Davier, A. A., & Wilson, C. (2005). *A didactic approach to the use of IRT true-score equating model* (Research Report No. RR-05-26). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2007a). *Hierarchical general diagnostic models* (Research Report No. RR-07-19). Princeton, NJ: Educational Testing Service.

von Davier, M. (2007b). *Mixture distribution diagnostic models* (Research Report No. RR-07-32). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61,* 287–307.

von Davier, M. (2008b). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 255–274). Charlotte, NC: Information Age Publishing.

von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York, NY: Springer Science & Business Media.

von Davier, M., DiBello, L., & Yamamoto, K. (2006). *Reporting test outcomes using models for cognitive diagnosis* (Research Report No. RR-06-28). Princeton, NJ: Educational Testing Service.

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151–174). Cambridge, MA: Hogrefe & Huber.

von Davier, M. & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68,* 213–228.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 643–661). Amsterdam, The Netherlands: Elsevier.

von Davier, M., & Sinharay, S. (2004). *Application of the stochastic EM method to latent regression models* (Research Report No. RR-04-34). Princeton, NJ: Educational Testing Service.

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics, 32,* 233–251.

von Davier, M., & Sinharay, S. (2009). *Stochastic approximation methods for latent regression item response models* (Research Report No. RR-09-09). Princeton, NJ: Educational Testing Service.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35,* 174–193.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao, & S. Sinharay S. (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam, The Netherlands: Elsevier.

von Davier, M., & von Davier, A. A. (2004). *A unified approach to IRT scale linking and scale transformation* (Research Report No. RR-04-09). Princeton, NJ: Educational Testing Service.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformation. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*, 115–124.

von Davier, M., & von Davier, A. A. (2010). A general model for IRT scale linking and scale transformation. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 225–242). New York, NY: Springer.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.

von Davier, M., & Yamamoto, K. (2003). *Partially observed mixtures of IRT models: An extension of the generalized partial credit model* (Research Report No. RR-03-22). Princeton, NJ: Educational Testing Service.

von Davier, M., & Yamamoto, K. (2007). Mixture distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 99–115). New York, NY: Springer.

Wainer, H. (1981). *Some standard errors in item response theory* (Research Report No. RR-81-36). Princeton, NJ: Educational Testing Service.

Wainer, H. (1983). On item response theory and computerized adaptive tests: The coming technological revolution in testing. *Journal of College Admissions*, *28*, 9–16.

Wainer, H. (1990). Introduction and history. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), C*omputerized adaptive testing: A primer* (pp. 1–21). Hillsdale, NJ: Erlbaum.

Wainer, H. (2000). Introduction and history. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 1–21). Mahwah, NJ: Erlbaum.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (2000). *Computer adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (1990). *Computer adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (pp. 233–270). Hillsdale, NJ: Erlbaum.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). Future challenges. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 231–269). Mahwah, NJ: Erlbaum.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 271–299). Mahwah, NJ: Erlbaum.

Wainer, H., & Lewis, C. (1989). *Toward a psychometrics for testlets* (Research Report No. RR-89-29). Princeton, NJ: Educational Testing Service.

Wainer, H., & Mislevy, R. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, D. R., Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (pp. 65–102). Hillsdale, NJ: Erlbaum.

Wainer, H., & Mislevy, R. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 61–100). Mahwah, NJ: Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991a). *Differential testlet functioning: Definitions and detection* (Research Bulletin No. RB-91-21). Princeton, NJ: Educational Testing Service.

Wainer, H., Sireci, S. G., & Thissen, D. (1991b) Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Wainer, H., & Thissen, D. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

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

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203–220.

Wainer, H., & Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL* (Research Report No. RR-01-09). Princeton, NJ: Educational Testing Service.

Wainer, H., Wang, X.-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees choice? *Journal of Educational Measurement, 31,* 183–199.

Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing theory* (Research Bulletin No. RB-76-08). Princeton, NJ: Educational Testing Service.

Wang, X., Bradlow, E. T., & Wainer, H. (2001). *User's guide for SCORIGHT (version 1.2): A computer program for scoring tests built of testlets* (Research Report No. RR-01-06). Princeton, NJ: Educational Testing Service.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). *A general Bayesian model for testlets: Theory and application* (Research Report No. RR-02-02). Princeton, NJ: Educational Testing Service.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *User's guide for SCORIGHT (Version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis* (Research Report No. RR-04-49). Princeton, NJ: Educational Testing Service.

Wendler, C. L. W., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445–467). Mahwah, NJ: Erlbaum.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver, Canada: Educational Research Institute of British Columbia.

Wingersky, M. S. (1987). *One-stage LOGIST* (Research Report No. RR-87-45). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Xu, X. (2007). *Monotone properties of a general diagnostic model* (Research Report No. RR-07-25). Princeton, NJ: Educational Testing Service.

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*, 121–137.

Xu, X., Douglas, J., & Lee, Y.-S. (2010). Linking with nonparametric IRT models. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 243–258). New York, NY: Springer.

Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution* (Research Report No. RR-11-41). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2006). *Cognitive diagnostics for NAEP proficiency data* (Research Report No. RR-06-08). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008a). *Comparing multiple-group multinomial log-linear models for multidimensional skill distributions in the general diagnostic model* (Research Report No. RR-08-35). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008b). *Fitting the structured general diagnostic model to NAEP* data (Research Report No. RR-08-27). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008c). *Linking for the general diagnostic model* (Research Report No. RR-08-08). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008d). Linking for the general diagnostic model. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments 1*, 97–111.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1995*). Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (Research Report No. RR-95-02). Princeton, NJ: Educational Testing Service.

Yamamoto, K., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait class models in the social sciences* (pp. 89–99). New York, NY: Waxmann.

Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 275–295). Hillsdale NJ: Erlbaum.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17,* 155–173.

Yan, D., Almond, R. G., & Mislevy, R. J. (2004). *A comparison of two models for cognitive diagnosis* (Research Report No. RR-04-02). Princeton, NJ: Educational Testing Service.

Yan, D., Lewis, C., Stocking, M. L. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics, 29,* 293–316.

• Yen, W. M. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 123–141). Vancouver, Canada: Educational Research Institute of British Columbia.

Yen, W. M., & Fitzpatrick, R. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger Publishers.

Zhang, J. (2004a). *Comparison of unidimensional and multidimensional approaches to IRT parameter estimation* (Research Report No. RR-04-44). Princeton, NJ: Educational Testing Service.

Zhang, J. (2004b). *Conditional covariance theory and DETECT for polytomous items* (Research Report No. RR-04-50). Princeton, NJ: Educational Testing Service.

Zhang, J. (2005a). *Bias correction for the maximum likelihood estimate of ability* (Research Report No. RR-05-15). Princeton, NJ: Educational Testing Service.

Zhang, J. (2005b). *Estimating multidimensional item response models with mixed structure* (Research Report No. RR-05-04). Princeton, NJ: Educational Testing Service.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72,* 69–91.

Zhang, J., & Lu, T. (2007). *Refinement of a bias-correction procedure for the weighted likelihood estimator of ability* (Research Report No. RR-07-23). Princeton, NJ: Educational Testing Service.

Zhang, J., & Stout, W. (1997). On Holland's Dutch identity conjecture. *Psychometrika, 62,* 375–392.

Zhang, J. & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129–152.

Zhang, J. & Stout, W. (1999b). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249.

Zwick, R. J. (1986). *Assessment of the dimensionality of NAEP year 15 reading data* (Research Report No. RR-86-04). Princeton, NJ: Educational Testing Service.

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

Zwick, R. J. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Zwick, R. J. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10–16.

Zwick, R. J., Thayer, D. T., & Wingersky, M. S. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32,* 341–363.

Notes

[1] Boldface and full spelling of an individual's name indicates an ETS staff member.

[2] Green stated that Tucker was at Princeton and ETS from 1944 to 1960; as head of statistical analysis at ETS, Tucker was responsible for setting up the statistical procedures for test and item analysis, as well as equating.

[3] As was common practice for many years at ETS, Green's (1950a) and Lord's (1952b) research bulletins (RBs) later resulted in journal articles (Green, 1951a; Lord 1953). In cases where two references are the same, or at least based on the same study (RBs and reports were often more detailed than the journal article), we will cite them together, separated by a slash (as in "1952b/1953"). Also, these research bulletins and journal articles by Green and Lord are based on their Ph.D. theses at Princeton University, both presented in 1951.

[4] Lord (1980a, p. 19) attributes the term *local independence* to Lazarsfeld (1950) and mentions that Lazarsfeld used the term *trace line* for a curve like the ICC. Rasch (1960) makes no mention of the earlier works referred to by Lord so we have to assume he was unaware of them or felt they were not relevant to his research direction.

[5] Samejima produced this work while at ETS. She later developed her GR models more fully while holding university positions. Her ETS Research Bulletin (1968) was also published as a Psychometric Monograph (1969).

[6] In this case, the available (on the ETS website) copy of the (1967) research bulletin is a copy of the (1968a) journal article.

[7] Developed in 1991 (as cited in Yen & Fitzpatrick, 2006), about the same time as Muraki was developing the GPC model.

[8] Unfolding models are proximity IRT models developed for assessments with binary disagree-agree or graded disagree-agree responses. Responses on these assessments are not necessarily cumulative and one cannot assume that higher levels of the latent trait will, lead to higher item scores and thus to higher total test scores. Unfolding models predict item scores and total scores on the basis of the distances between the test taker and each item on the latent continuum (Roberts, n.d.).

[9] While these authors were not ETS staff members, this report was completed under the auspices of the External Diagnostic Research Team, supported by ETS.

[10] The bullet symbol in the reference list (•) indicates work that was not performed at ETS.

Reports in the ETS R&D Scientific and Policy Contributions Series

Reports in the ETS R&D Scientific and Policy Contributions Series document the contributions made by the research program at Educational Testing Service since the founding of the organization in 1947.

Evaluating Educational Programs
by Samuel Ball (2011)
ETS R&D Scientific and Policy Contributions Series No. SPC-11-01

Modeling Change in Large-Scale Longitudinal Studies of Educational Growth: Four Decades of Contributions to the Assessment of Educational Growth
by Donald A. Rock (2012)
ETS R&D Scientific and Policy Contributions Series No. SPC-12-01

Understanding the Impact of Special Preparation for Admissions Tests
by Donald E. Powers (2012)
ETS R&D Scientific and Policy Contributions Series No. SPC-12-02

ETS Research on Cognitive, Personality, and Social Psychology: I
by Lawrence J. Stricker (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-01

Contributions of a Nonprofit Educational Measurement Organization to Education Policy Research
by Richard J. Coley, Margaret E. Goertz, and Gita Z. Wilder (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-02

ETS Psychometric Contributions: Focus on Test Scores
by Tim Moses (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-03

ETS Contributions to the Quantitative Assessment of Item, Test, and Score Fairness
by Neil J. Dorans (2013)
ETS R&D Scientific and Policy Contributions Series No. SPC-13-04