# ITEM RESPONSE THEORY FOR PSYCHOLOGISTS

SUSAN E. EMBRETSON
STEVEN P. REISE

# *Item Response Theory for Psychologists*

## MULTIVARIATE APPLICATIONS BOOKS SERIES

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

There are currently four books in the series:

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

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

# *Item Response Theory for Psychologists*

Susan E. Embretson
*University of Kansas*

Steven P. Reise
*University of California*


LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2000 Mahwah, New Jersey London

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

Cover design by Kathryn Houghtaling Lacey

**Library of Congress Cataloging-in-Publication Data**

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

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

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

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

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

# *Contents*

# *Preface*

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

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

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

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

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

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

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

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

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

## ACKNOWLEDGMENTS

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

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

*—Susan E. Embretson*
*—Steven P. Reise*

# I

# *INTRODUCTION*

# 1

# *Introduction*

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

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

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

## SCOPE OF IRT APPLICATIONS

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

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

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

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

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

## HISTORY OF IRT

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

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

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

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

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

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

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

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

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

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

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

## THE ORGANIZATION OF THIS BOOK

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

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

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

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

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

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

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

# II

# *ITEM RESPONSE THEORY PRINCIPLES: SOME CONTRASTS AND COMPARISONS*

# 2

# *The New Rules of Measurement*

Classical test theory (CTT) has been the mainstay of psychological test development for most of the 20th century. Gulliksen's (1950) classic book, which remains in print, is often cited as the defining volume. However, CTT is much older. Many procedures were pioneered by Spearman (1907, 1913). CTT has defined the standard for test development, beginning with the initial explosion of testing in the 1930s.

However, since Lord and Novick's (1968) classic book introduced model-based measurement, a quiet revolution has occurred in test theory. Item response theory (IRT) has rapidly become mainstream as a basis for psychological measurement. IRT, also known as latent trait theory, is model-based measurement in which trait level estimates depend on both persons' responses and on the properties of the items that were administered. Many new or revised tests, particularly ability tests, have been developed from IRT principles. Yet, because most test users are unfamiliar with IRT, test manuals mention its application only in passing or in a technical appendix. Thus test users are largely unaware that the psychometric basis of testing has changed.

Initially, IRT appealed to U.S. test developers because it solved many practical testing problems, such as equating different test forms (see Lord, 1980). More recently, the promise of IRT for substantive issues in psychology has become apparent. Score interpretations may now be related to underlying skills through the conjoint measurement properties of IRT (see Rule 10). Furthermore, the justifiable measurement scale properties for IRT can have significant impact on inferential statistics about group differences (Embretson, 1996a; Maxwell & Delaney, 1985), as well as on test score comparisons within or between persons.

Most psychologists' knowledge of the "rules of measurement" is based on CTT. Test theory is included in the curriculum for applied psychologists and for many theoretically inclined psychologists. In some graduate programs, CTT is presented in a separate course, which is required for applied psychologists and elective for other areas. In other graduate programs, CTT is elaborated in testing methods for courses for clinical, counseling, industrial, and school psychologists.

To provide continuity between the new test theory and the old test theory, Lord and Novick (1968) derived many CTT principles from IRT. On the surface, this seems to be good news for the busy psychologist who knows CTT but not IRT. The existence of the derivations seemingly suggests that the rules of measurement, although rooted in a more sophisticated body of axioms, remain unchanged.

Yet, in the new model-based version of test theory, IRT, some well-known rules of measurement no longer apply. In fact, the new rules of measurement are fundamentally different from the old rules. Many old rules, in fact, must be revised, generalized, or abandoned altogether.

This chapter contrasts several old rules of measurement to the corresponding new rules of measurement to illustrate the depth of the differences between CTT and IRT. These differences are described in their most extreme form in this chapter. Although many extensions of CTT were developed to handle some of the problems noted in this chapter, their application has been limited. We believe that for the many hundreds of small-volume psychological tests, the old rules are valid.

## A COMPARISON OF MEASUREMENT RULES

Several old rules of measurement may be gleaned from the principles of CTT or its common extension. Other old rules are implicit in many applied test development procedures. Table 2.1 shows 10 old rules that are reviewed here. The old rules are followed by 10 corresponding new rules, which obviously conflict with the old rules.

We argue that the old rules represent common knowledge or practice among psychologists. These rules have guided the development of many, but certainly not all, published psychological tests. Obvious exceptions include some selection and admissions tests that were developed by large-scale testing corporations and the military. In these cases, non-IRT procedures were developed to circumvent the limitations of some old rules. For example, nonlinear test equating (see Holland & Rubin, 1982) was developed to handle the limitation of Old Rule 3. Also, population-free item indices, such as the delta index used by ETS (see Gulliksen, 1950, p. 368), was developed to counter Old Rule 4, respectively. Last, pro-

TABLE 2.1
Some "Rules" of Measurement

| | *The Old Rules* |
|---|---|
| Rule 1. | The standard error of measurement applies to all scores in a particular population. |
| Rule 2. | Longer tests are more reliable than shorter tests. |
| Rule 3. | Comparing test scores across multiple forms is optimal when the forms are parallel. |
| Rule 4. | Unbiased estimates of item properties depend on having representative samples. |
| Rule 5. | Test scores obtain meaning by comparing their position in a norm group. |
| Rule 6. | Interval scale properties are achieved by obtaining normal score distributions. |
| Rule 7. | Mixed item formats leads to unbalanced impact on test total scores. |
| Rule 8. | Change scores cannot be meaningfully compared when initial score levels differ. |
| Rule 9. | Factor analysis on binary items produces artifacts rather than factors. |
| Rule 10. | Item stimulus features are unimportant compared to psychometric properties. |
| | *The New Rules* |
| Rule 1. | The standard error of measurement differs across scores (or response patterns), but generalizes across populations. |
| Rule 2. | Shorter tests can be more reliable than longer tests. |
| Rule 3. | Comparing test scores across multiple forms is optimal when test difficulty levels vary between persons. |
| Rule 4. | Unbiased estimates of item properties may be obtained from unrepresentative samples. |
| Rule 5. | Test scores have meaning when they are compared for distance from items. |
| Rule 6. | Interval scale properties are achieved by applying justifiable measurement models. |
| Rule 7. | Mixed item formats can yield optimal test scores. |
| Rule 8. | Change scores can be meaningfully compared when initial score levels differ. |
| Rule 9. | Factor analysis on raw item data yields a full information factor analysis. |
| Rule 10. | Item stimulus features can be directly related to psychometric properties. |

cedures have been developed to estimate measurement errors at specific score levels (see Feldt & Brennan, 1989) to counter Old Rule 1. However, these techniques are not well known outside large-scale testing programs, hence they are not routinely applied in the development of psychological tests. Thus the old rules characterize substantial practice in test development.

## Rule 1: The Standard Error of Measurement

*Old Rule 1: The standard error of measurement applies to all scores in a particular population.*

*New Rule 1: The standard error of measurement differs across scores but generalizes across populations.*

These two rules concern the properties of the standard error of measurement. The standard error of measurement describes expected score fluctuations due to error. Not only is the standard error of measurement basic to describing the psychometric quality of a test, it is also critical to individual score interpretations. The confidence intervals defined by the standard error can guide score interpretations in several ways. Differences between two scores, for example, may be interpreted as not significant if their confidence interval bands overlap.

New Rule 1 conflicts with Old Rule 1 in two ways. First, the rules differ in whether the standard error of measurement is constant or variable among scores in the same population. Old Rule 1 specifies constancy, whereas New Rule 1 specifies variability. Second, the rules differ in whether the standard error of measurement is specific or general across populations. Old Rule 1 is population-specific, whereas New Rule 1 is population-general.

The basis of the standard error of measurement differs substantially between CTT and IRT. Thus we elaborate these differences somewhat more here. A more complete elaboration of the IRT standard error of measurement is presented in chapter 7.

In CTT, the standard error of measurement is computed as the square root of 1 minus reliability $(1 - r_{tt})^{1/2}$, times the standard deviation of the test $\sigma$ as follows:

$$SE_{\text{Msmt}} = (1 - r_{tt})^{1/2}\sigma \tag{2.1}$$

Confidence intervals are constructed for individual scores under the assumption that measurement error is distributed normally and equally for all score levels.

To illustrate the old rule, item response data were simulated for 3,000 examinees on a 30-item test with a normal difficulty range. The examinees were sampled from a population with a standard normal distribution of trait scores (see Embretson, 1994b, for details). In the upper panel of Fig. 2.1, classical true scores (shown as standard scores) are regressed on raw test scores. A 68% confidence interval, using a standard error of .32 (e.g., from Cronbach's alpha index of internal consistency), is shown by the dotted lines.

Two important points about the old rule may be gleaned from the upper panel of Fig. 2.1. First, the estimated true score is a standard score that is derived as a *linear* transformation of raw score, as noted by the linear regression. Second, the confidence intervals are also represented as straight lines for all scores because the same confidence interval applies to each score. In CTT, both the transformation of raw score to true score and the standard error apply to a particular population because their estimation

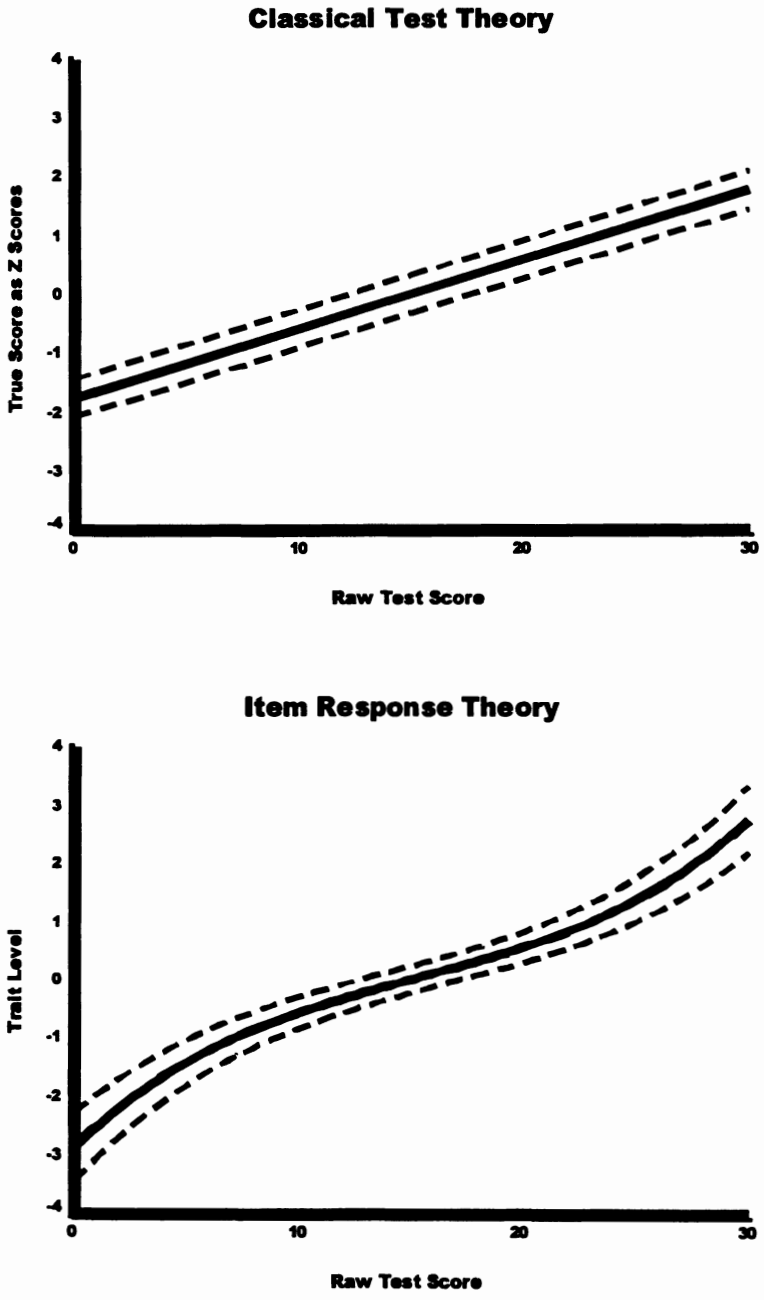


FIG. 2.1. Regression of true score on raw score.

depends on population statistics. That is, the standard score conversion requires estimating the raw score mean and standard deviation for a population, whereas the standard error requires estimating both the variance and reliability.

To illustrate the new rule, trait scores were estimated using the Rasch IRT model on the simulation data that were described earlier. In the lower panel of Fig. 2.1, IRT trait level scores are regressed on raw scores. The lower panel differs from the upper panel of Fig. 2.1 in two important ways: First, the relationship between trait score and raw score is *nonlinear;* second, the confidence interval band becomes increasingly wide for extreme scores. Unlike CTT, neither the trait score estimates nor their corresponding standard errors depend on population distributions. In IRT models, trait scores are estimated separately for each score or response pattern, controlling for the characteristics (e.g., difficulty) of the items that were administered. Standard errors are smallest when the items are optimally appropriate for a particular trait score level and when item discriminations are high. The details of IRT standard errors are provided in later chapters.

The standard errors from IRT may be averaged to provide a summary index for a population. A composite reliability (see Andrich, 1988b) may be computed for the group by comparing the averaged squared standard errors, $\sigma_\theta^2$, to the trait score variance, $\sigma^2$, as follows:

$$r'_{tt} = 1 - \frac{\sigma_\theta^2}{\sigma^2} \tag{2.2}$$

Obviously, the smaller the standard error at each trait score level, the higher the reliability. In the large normal population of simulated examinees, with a test that was appropriate for its trait level (see Embretson, 1995c), the *average* standard error, across examinees, was .32, the same as the uniform standard error from CTT in the first analysis. However, these standard errors will be similar only in limited circumstances. For example, if test difficulty was not well matched to the sample or if the distribution was not normal, differences would be expected.

### Rule 2: Test Length and Reliability

*Old Rule 2: Longer tests are more reliable than shorter tests.*
*New Rule 2: Shorter tests can be more reliable than longer tests.*

These two rules contrast directly, and even surprisingly. It is axiomatic in CTT that longer tests are more reliable. In fact, this principle is represented by an equation in CTT; namely, the Spearman-Brown prophesy formula. Specifically, if a test is lengthened by a factor of $n$ parallel parts,

true variance increases more rapidly than error variance (Guilford, 1954, presents the classic proof). If $r_{tt}$ is the reliability of the original test, the reliability of the lengthened test $r_{nn}$ may be anticipated as follows:

$$r_{nn} = \frac{n\, r_{tt}}{1 + (n-1) r_{tt}} \tag{2.3}$$

Equation 2.3 may also be applied to shortened tests. That is, if a test with a reliability of .86 is shortened to two-thirds of the original length ($n = .667$), then the anticipated reliability of the shorter test is .80. Figure 2.2 shows the effect of doubling, tripling, and so on, a test with an initial reliability of .70 with parallel items.

The new rule counters the old rule by asserting that short tests can be more reliable. Figure 2.3 shows the standard error of measurement at various levels of trait score for various lengths and types of tests based on the simulation data. Item discrimination was held constant in all analyses. All results in the lower panel are based on IRT. The two fixed content tests, of 20 and 30 items, respectively, show the characteristic IRT pattern of higher measurement errors for extreme scores. Also notice that the standard errors from the 30-item test are smaller than those from the 20-item test at all trait levels. This pattern is consistent with the *old rule.*

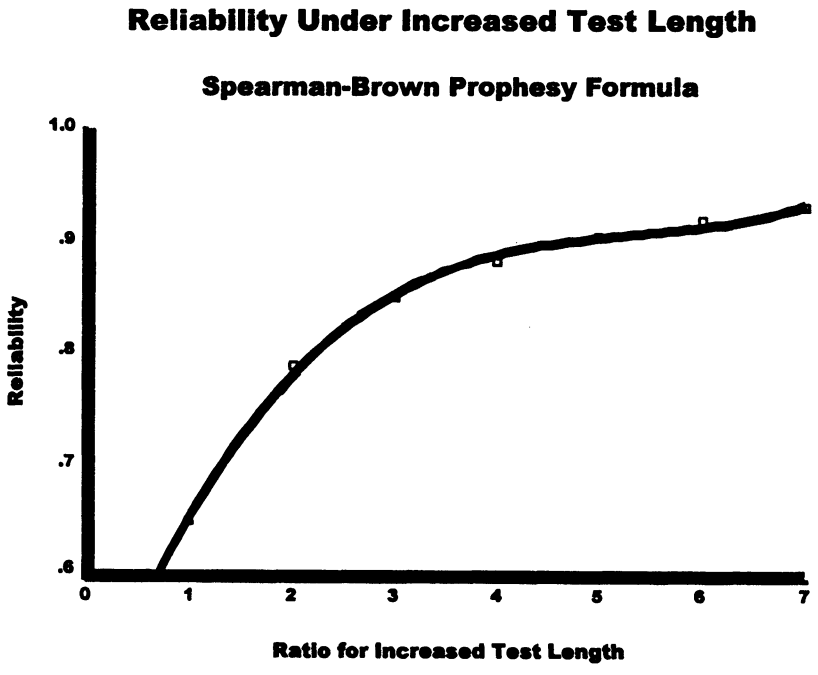


FIG. 2.2. Impact of test length in classical test theory.

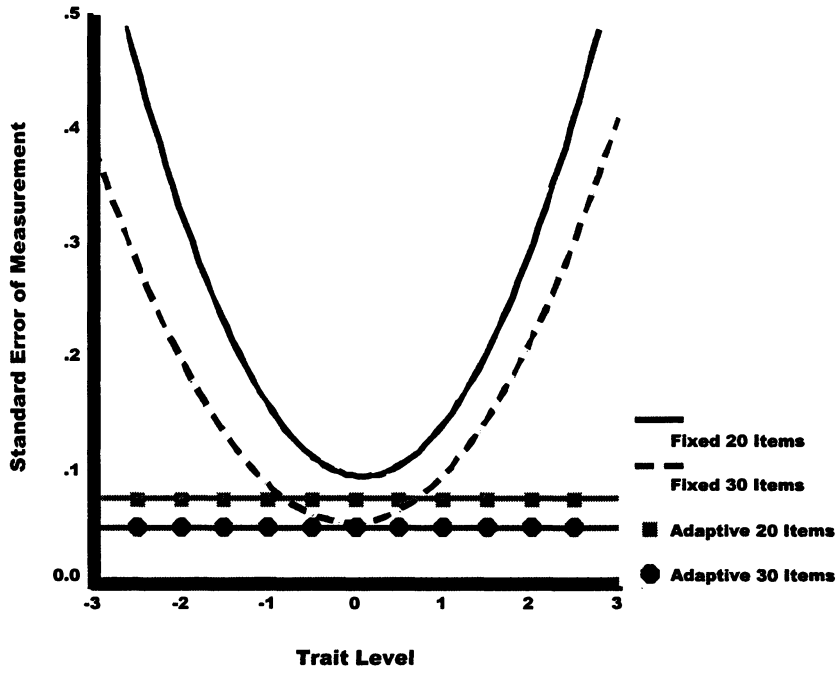


FIG. 2.3. Impact of trait level and test length on measurement error in IRT.

The standard errors from two adaptive tests also are shown on the lower panel of Fig. 2.3. In adaptive tests, items are selected for each examinee to be most appropriate for their ability level. Thus, examinees at different ability levels will be administered different items. Given sufficient coverage of item difficulty in the item bank, equal standard errors can be obtained for each trait level. In the simulation study, Fig. 2.3 shows that equal standard errors were obtained across trait levels.

The new rule is illustrated by comparing the standard errors between traditional fixed content tests and adaptive tests. Notice that the standard error from the 20-item adaptive test is *lower* for most trait levels than from the 30-item fixed content test. This is a typical pattern for adaptive testing. The implication, of course, is that the shorter test yields less measurement error. Thus, a more reliable "test" has been developed at a shorter length with items of the same quality (i.e., item discriminations are constant). Thus, a composite reliability across trait levels as in Eq. 2.1, would show the shorter (adaptive) test as more reliable than the longer normal range test.

In fairness to CTT, it should be noted that an assumption underlying the Spearman-Brown prophesy formula is that the test is lengthened with

parallel parts. An adaptive test, by its nature, fails to meet this assumption. The item difficulties vary substantially in the adaptive tests. However, the point made here is that the old rule about test length and reliability conflicts sharply with current adaptive testing practice, which is based on the new rule from IRT.

### Rule 3: Interchangeable Test Forms

*Old Rule 3: Comparing test scores across multiple forms is optimal when test forms are parallel.*

*New Rule 3: Comparing test scores across multiple forms is optimal when test difficulty levels vary between persons.*

When examinees receive different test forms, some type of equating is needed before their scores can be compared. Traditionally, CTT relied on test form parallelism to equate scores. Gulliksen (1950) defined strict conditions for test parallelism in his exposition of CTT. The conditions included the equality of means and variances across test forms, as well as equality of covariances with external variables. If test forms meet Gulliksen's statistical conditions for parallelism, then scores may be regarded as comparable across forms.

Practically, however, test form parallelism cannot be met. Test form means and variances often differ somewhat. Furthermore, often score comparisons between rather different tests sometimes are desired. Thus, substantial effort has been devoted to procedures for test equating (see Angoff, 1982; Holland & Rubin, 1982). Comparable scores between test forms are established by techniques such as linear equating and equipercentile equating. In linear equating, for example, scores on one test form are regressed on the other test form. The regression line can then be used to find the comparable scores from one test form to the other.

Although the various equating methods can be applied when the test forms that have different means, variances, and reliabilities; equating error is influenced by differences between the test forms. Equating error is especially influenced by differences in test difficulty level (see Peterson, Marco, & Stewart, 1982). Thus, although scores can be linked between test forms, equating error can be problematic. Test forms with high reliabilities and similar score distributions will be most adequately equated.

The adverse effect of test difficulty differences on equating scores may be illustrated by considering some further results from the simulation study. In the simulation, 3,000 examinees were given two test forms containing the same item discriminations but substantially different item difficulties. Item responses for the easy test and the hard test were generated.

To illustrate the equating problem, consider the regression of the scores from the easy test on scores from a hard test shown in Fig. 2.4. The spread of scores on the easy test is shown at each score on the hard test. The best-fit linear regression between the tests yielded a squared multiple correlation of .67. This correlation is not very high for comparing scores across forms.

Thus the data shown in Fig. 2.4 have two problems for equating. First, a nonlinear regression is needed to fully describe score correspondence. For example, Fig. 2.4 shows that with linear equating, the easy test scores are underestimated at some score levels and overestimated at others. Nonlinear equating is significantly better in this case. The best-fit cubic relationship yielded a squared multiple correlation of .84, which leads to a lower equating error. Second, however, equating error will be substantial even for nonlinear equating. The variance of easy test scores at low scores on the hard test is quite large. For example, for an observed score of zero on the hard test, examinees scores on the easy test range from 0 to 20. The hard test simply does not have the floor to distinguish these examinees, so equating cannot be very satisfactory. Similarly, high scores on the hard test are associated with the same (perfect) score on the easy test. Thus no method of equating will match scores appropriately.

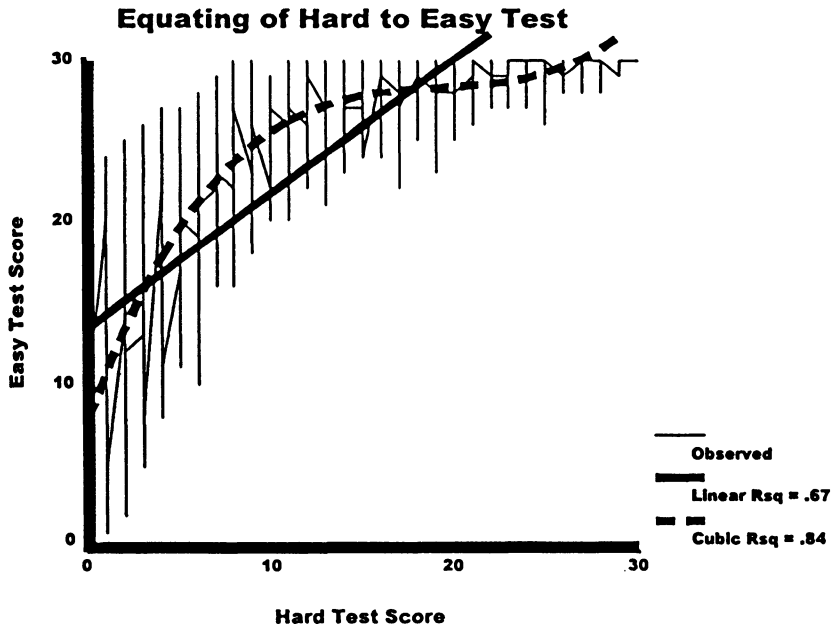


FIG. 2.4. Regression of easy test scores on hard test scores.

Because the scores shown in Fig. 2.4 are based on simulated data, the true trait levels are known. Thus the reliability index for the classical test scores could be computed directly by correlating raw scores with true scores. The upper panel of Fig. 2.5 shows the relationship of the two classical test forms with true trait level (i.e., the generating value in the simulation). Two points are worth noting in the upper panel. First, the relationship of scores on both tests are nonlinearly related to true trait level. That is, the relative distances between true scores are not reflected by the relative distances between raw scores on either test. Second, the squared correlation between true score and raw score is .8666 and .8736, respectively, for the hard test and the easy test. Although this value is high, it is rather low for simulation data based on a unidimensional trait underlying the items.

The lower panel of Fig. 2.5 shows the relationship of true trait level to estimated trait level from a simulated adaptive test of 30 items. In the adaptive test, examinees receive different subsets of items, which vary substantially in item difficulty. In fact, item difficulty is selected to be optimally appropriate for each examinee in an adaptive test. For persons with high trait levels, difficult items will be selected, whereas for persons with low trait levels, easier items will be selected. Trait level is estimated separately for each person using an IRT model that controls for differences in item difficulty. Notice that the squared correlation of the true trait level with the adaptive test score is much higher than for the two classical tests ($r^2 = .9695$).

In summary, Fig. 2.5 shows the new rule clearly. Better estimation of true trait level is obtained by *nonparallel* test forms. That is, each adaptive test is a separate test form that differs substantially—and deliberately—in difficulty level from other forms. The correlation of estimated trait level with true trait level is substantially higher than the correlations for the classic tests in the upper panel of Fig. 2.5. An explanation of how item difficulty influences measurement error is provided in several chapters, especially chapters 3 and 7.

### Rule 4: Unbiased Assessment of Item Properties

*Old Rule 4: Unbiased assessment of item properties depends on having representative samples.*

*New Rule 4: Unbiased estimates of item properties may be obtained from unrepresentative samples.*

The CTT statistic for item difficulty is *p*-value, which is computed as the proportion passing. The CTT statistic for item discrimination is item-total correlation (e.g., biserial correlation). Both statistics can differ substantially

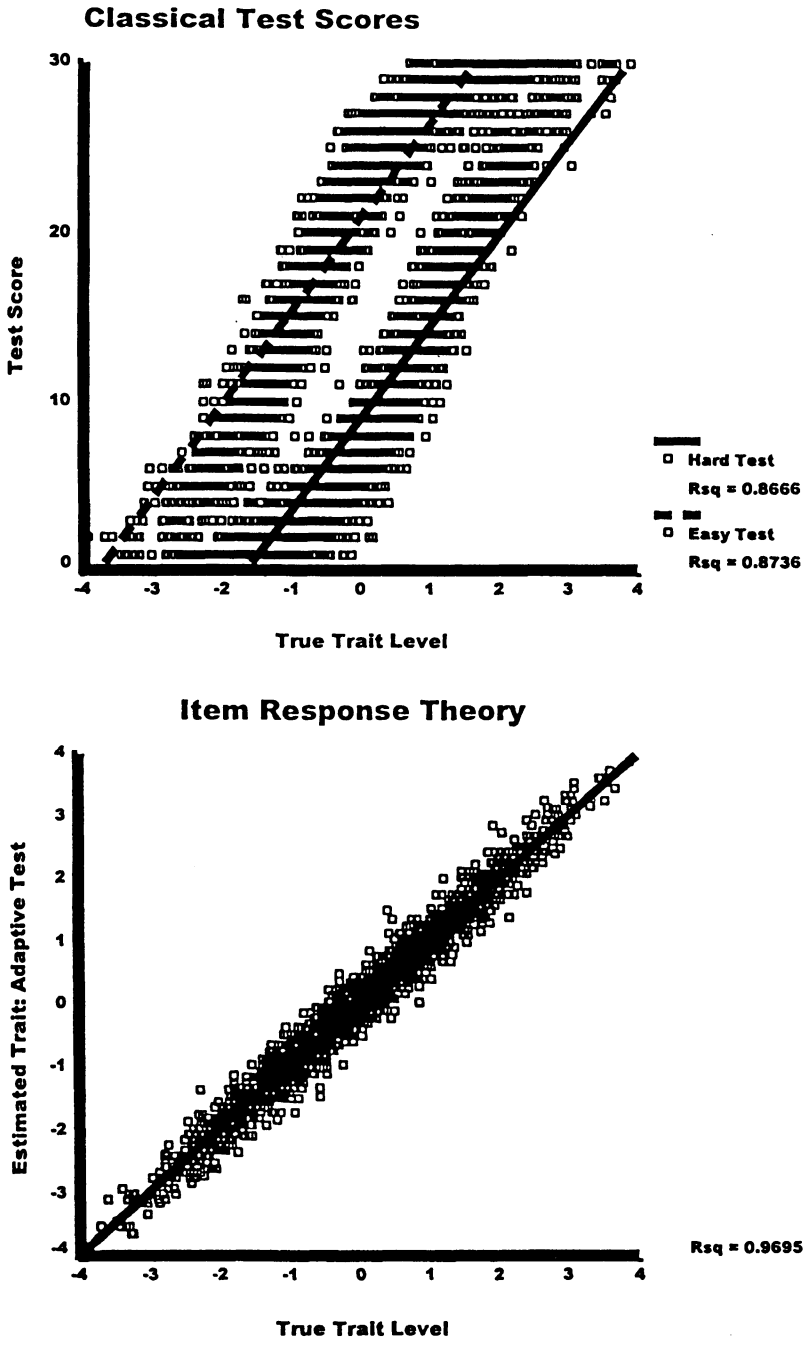


FIG. 2.5. Regression of test scores on true trait level.

across samples if computed from unrepresentative samples. To illustrate the effect, the 3,000 simulated examinees, described earlier, were split at the median into two extreme samples of 1,500 each. The high group had scores above the median, whereas the low group had scores below the median.

In the upper panel of Fig. 2.6, the estimated *p*-values for items are plotted by the two groups. A linear regression would indicate that the relative intervals between items is maintained. However, notice that the relationship between *p*-values, although generally monotonic, is not linear. The distances between items with high *p*-values is greater in the low ability sample, whereas the distances between items with low *p*-values is greater in the high ability sample. The correlation between *p*-values is only .800. The biserial correlations (not shown) of items with total score differed even more between groups.

The lower panel of Fig. 2.6 shows the item difficulty values that are obtained by a Rasch model scaling of the same data as shown in the upper panel. In the lower panel, unlike the upper panel, the correspondence of item difficulty values is quite close between the two extreme groups. The correlation between item difficulty values in the lower panel is .997.

### Rule 5: Establishing Meaningful Scale Scores

*Old Rule 5: Test scores obtain meaning by comparing their position in a norm group.*

*New Rule 5: Test scores obtain meaning by comparing their distance from items.*

The comparison standard for tests that stem from CTT is a relevant group of persons; namely, the norm group. The numerical basis of the comparison is order (i.e., position in the distribution). Standard scores are obtained by linear transformations of raw scores so that scores may be readily compared to positions in a normally distributed population of persons.

To illustrate the differences between the old and the new rule, Fig. 2.7 presents data from the Functional Independence Measurement (FIM) scale. The FIM is a behavioral report on activities of everyday living that often challenge elderly persons. The lower panel shows the items on the scale, ordered by difficulty. The FIM shows high internal consistency, which indicates a common order in the loss of these functions.

The upper panel of Fig. 2.7 shows the classical norm-referenced interpretation of FIM scores. In Fig. 2.7 is a histogram to show the frequencies of various FIM *z*-scores in a population of persons from age 80 to 89. Expectations from the normal distribution are overlaid. Here, the *z*-scores for

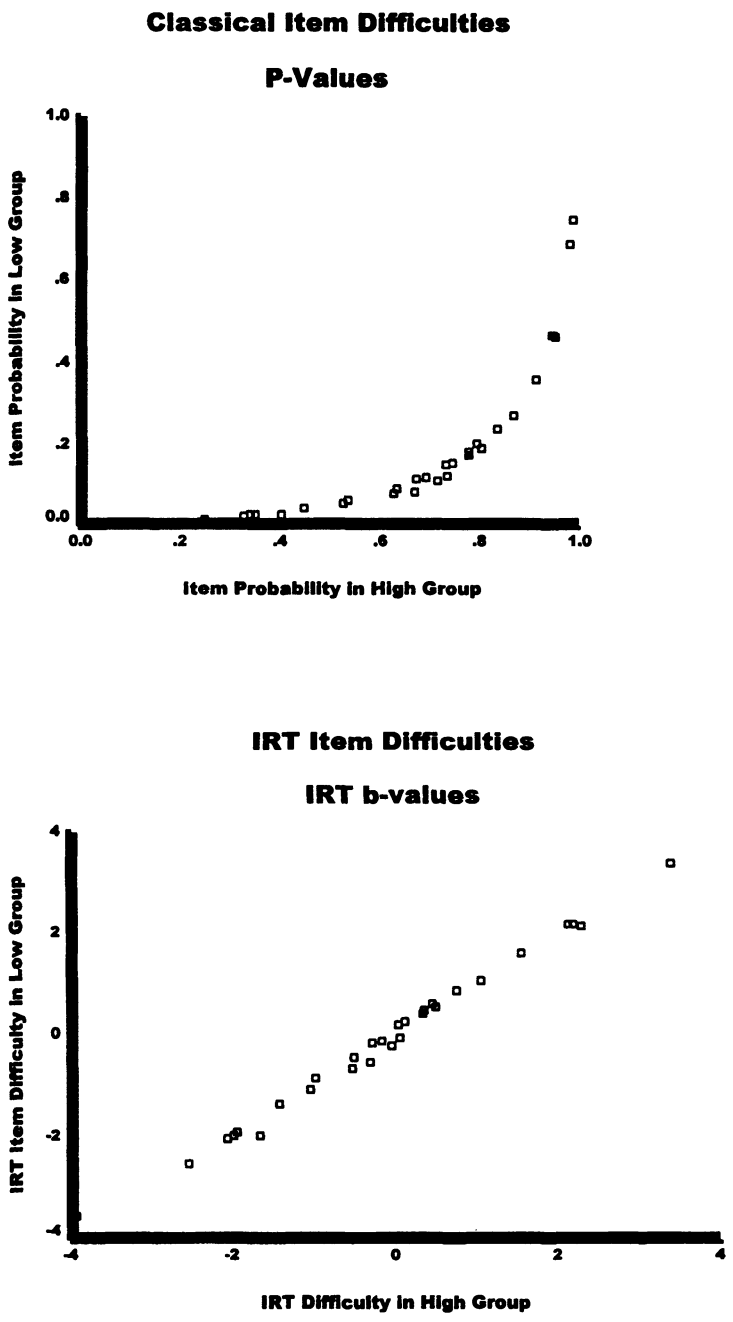


FIG. 2.6. Relationship between item difficulties obtained from two groups.

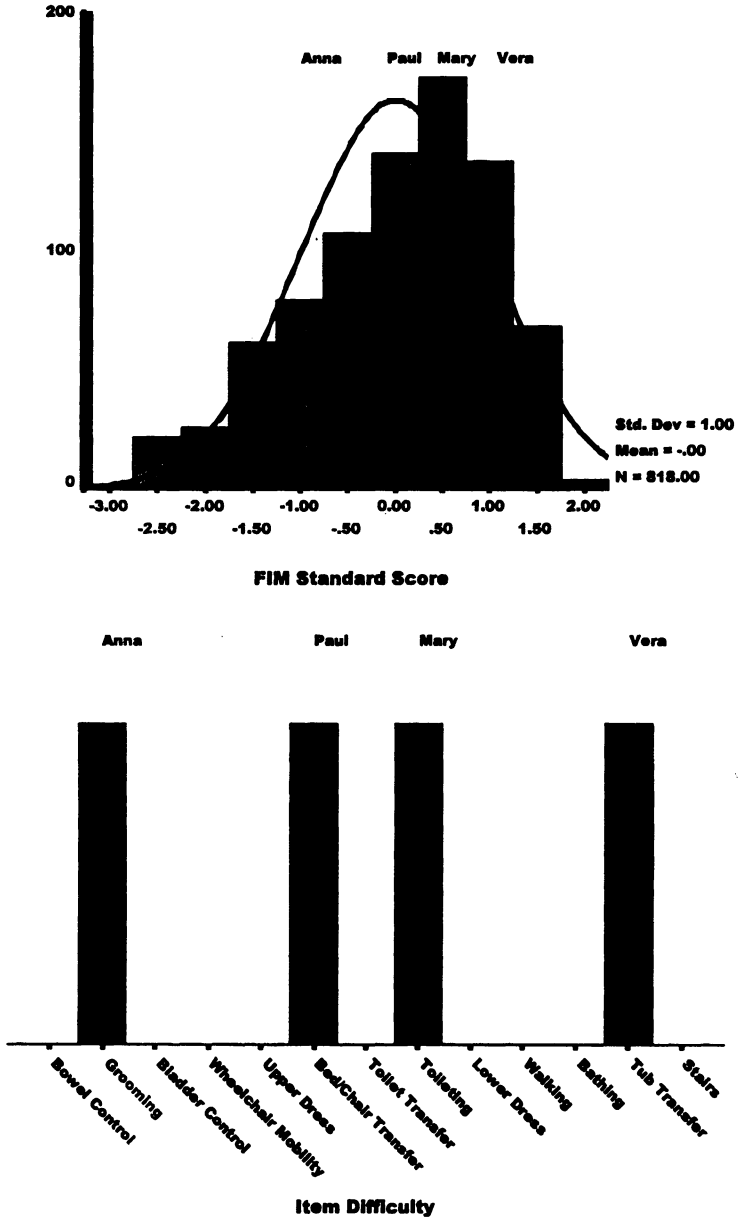


FIG. 2.7. Norm-referenced versus item-referenced meaning for the Functional Independence Measurement Scale.

four persons are projected into the distribution. Anna has a negative z-score, which indicates that her performance in everyday living skills is relatively low compared to her age group. Vera and Mary, in contrast, score relatively higher, while Paul is at the mean. Thus, we would conclude that Paul, Vera, and Mary are functioning relatively well for their age, whereas Anna is not. Score meaning, then, is determined primarily as location in a norm-referenced standard. Of course, as has been argued by proponents of criterion-referenced testing, norm-referenced scores have no meaning for what the person actually can do. That is, scores are not anchored to the skills represented by the items.

In IRT models, trait levels may be compared to items since persons and items are placed on a common scale. The difficulty of items is located on a continuum, typically ranging from –3 to +3, which is similar in magnitude to z-scores. The lower panel of Fig. 2.7 shows the relative location of the FIM items, ranging from most difficult (on the right) to most easy (on the left), as scaled by the Rasch model. The items on the right end are lost relatively early for elderly adults, while the items on the lower end are lost much later. The same four persons now are projected on the item scale. The black bars on the lower panel of Fig. 2.7 represents the location of the four persons with respect to the 13 FIM items.

The difference between a person's trait level and item difficulty has direct meaning for performance. If the person's trait score equals the item's difficulty, then the person is as likely to pass as to fail the item. Thus, items at the person's location in the scale have probability of .50 of being performed successfully. So the probability that Anna successfully completes "Grooming" is .50, while the probability that she completes the more difficult functions is lower than .50. For Paul, the probability that he completes "Bed/Chair Transfer" successfully is .50, but his probability is higher than .50 for all the lower activities, such as "Wheelchair Mobility" and "Grooming." Thus, meaning is referenced to the items.

It should be noted that IRT scaling of ability does not preclude linear transformations to standard scores so that norm-referenced meaning also may be obtained. The meaning of ability for item performance also may be retained by transforming item difficulties by a corresponding linear transformation.

### Rule 6: Establishing Scale Properties

*Old Rule 6: Interval scale properties are achieved by obtaining normal score distributions.*

*New Rule 6: Interval scale properties are achieved by applying justifiable measurement models.*

Although many test developers probably would not explicate Old Rule 6, the rule is implicit in routine test development procedures. Normal distributions are achieved in two ways. First, for many psychological tests, items are selected to yield normal distributions in a target population. Expositions of CTT typically show how normal score distributions may be achieved by selecting items with proportions passing around .50 (see Gulliksen, 1950). Second, for many other psychological tests, normal distributions are achieved by normalizing procedures. For example, intelligence tests are often scored as composites over several subtests. If the composites are not normally distributed in the target population, then either nonlinear transformations or percentile matching procedures may be applied. In percentile matching, the percentile rank of each raw score in the nonnormal distribution is calculated. Then, the normalized standard score is the z score in the normal distribution that corresponds to the percentile rank of the raw score. Normalizing by percentile matching is a *nonlinear* transformation that changes the relative distances between scores.

The relationship between score distributions and scale levels may be considered by revisiting the regression of easy test scores on hard test scores (i.e., Fig. 2.4). Scores were compressed on opposite ends of the continuum for the two tests. For the hard test, scores were compressed on the low end, while for the easy test, scores were compressed on the high end. Thus, the regression of hard test scores on easy test scores was nonlinear.

Score compression has two effects that are relevant to achieving interval level scales. First, the *relative distances* between scores are not constant under CTT scaling. The upper panel of Fig. 2.8 shows the best nonlinear equating of the easy test to the hard test. Consider a pair of low scores that are 5 points apart, at the score of 5 and 10 (see the reference lines). The distance between the equated scores is still 5 points, although of course now the scores are much higher (i.e., 20 and 25, respectively). Consider another pair of scores that are 5 points apart on the hard test; the scores at 20 and 25. Their equated scores on the easy test are not 5 points apart. Instead, their expected scores differ by less than one point. Failing to maintain constant distances between scores across tests implies that interval scale measurement is not achieved.

The lower panel of Fig. 2.8 shows how equal trait level differences in IRT led to the equal differences in performance expectations for an easy item versus a hard item. Constant differences in trait level imply constant differences in log odds for item success. Regardless of item difficulty level, the same difference in log odds is observed. Second, tests with scale compression have skewed score distributions in the target population. Figure 2.9 shows the frequency distributions for the easy and the hard tests, which are negatively and positively skewed, respectively.

If normal distributions are achieved, what happens to scale properties? Jones (1971) pointed out that under certain assumptions interval scale

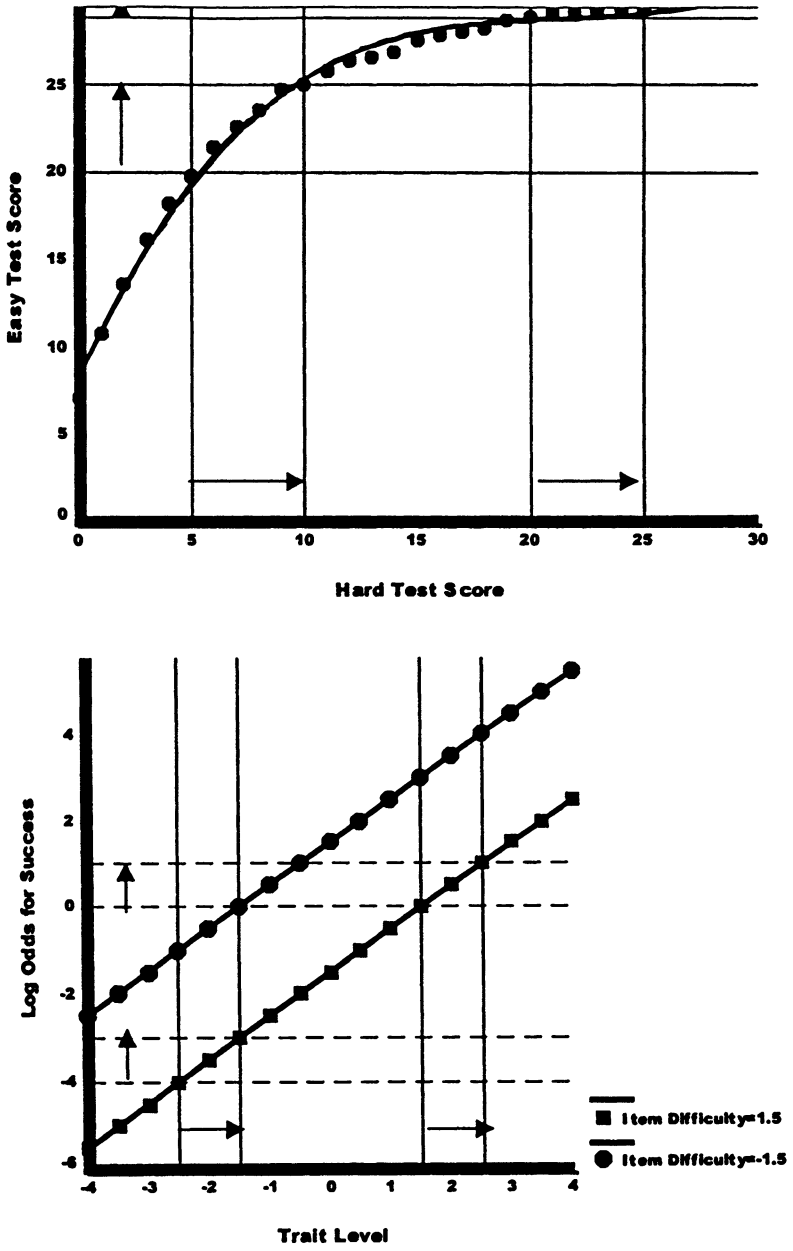


FIG. 2.8. The meaning of score distances in classical test theory versus IRT.