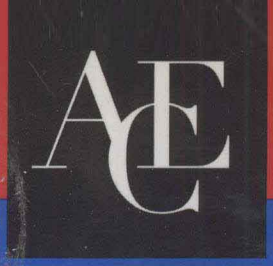

AMERICAN COUNCIL ON EDUCATION

Educational Measurement

FOURTH EDITION

EDITED BY

ROBERT L. BRENNAN

Sponsored Jointly by
National Council on Measurement in Education and
American Council on Education

ACE/PRAEGER SERIES ON HIGHER EDUCATION

EDUCATIONAL MEASUREMENT

Fourth Edition

Sponsored Jointly by
National Council on Measurement in Education and
American Council on Education

Edited by
Robert L. Brennan

AMERICAN COUNCIL ON EDUCATION
PRAEGER
Series on Higher Education

Library of Congress Cataloging-in-Publication Data

Educational measurement / sponsored jointly by National Council on Measurement in Education and American Council on Education ; edited by Robert L. Brennan. — 4th ed.
p. cm. — (ACE/Praeger series on higher education)
Includes index.
ISBN 0–275–98125–8
1. Educational tests and measurements. I. Brennan, Robert L. II. National Council on Measurement in Education. III. American Council on Education. IV. Series: American Council on Education/Praeger series on higher education.
LB3051.E266 2006
371.26—dc22 2006015706

British Library Cataloguing in Publication Data is available.

Library of Congress Catalog Card Number: 2006015706
ISBN: 0–275–98125–8

First published in 2006

Praeger Publishers, 88 Post Road West, Westport, CT 06881
An imprint of Greenwood Publishing Group, Inc.
www.praeger.com

Printed in the United States of America

The paper used in this book complies with the Permanent Paper Standard issued by the National Information Standards Organization (Z39.48–1984).

10 9 8 7 6 5 4 3 2 1

Foreword

Over 55 years ago the American Council on Education saw the need for and had the foresight to ask E. F. Lindquist to assemble a comprehensive, edited book related to educational measurement. The first edition of *Educational Measurement*, published in 1951, became a frequently used and highly regarded reference and classroom text on important and state-of-the-art topics in measurement. After a 20-year span, the second edition was published that was edited by Robert L. Thorndike. It contained chapters written by some of the most prominent names in the measurement community. The second edition both updated topics in the original and introduced new topics, and it became one of the most widely used and referenced books in the measurement field. In 1989, a third edition of this widely known and useful resource was published. The third edition was edited by Robert L. Linn, and in keeping with the previous editions, contained chapters written and reviewed by some of the leading measurement researchers. Now, 17 years later, yet another editor, Robert L. Brennan, has assembled widely known and respected researchers to write and review chapters that bring forth the most current thinking on traditional measurement topics and also to introduce new topics of importance in the educational measurement milieu.

This edition is the second time that the American Council on Education (ACE) and the National Council on Measurement in Education (NCME) have collaborated on the production of this book. These organizations also worked together to produce the third edition. Because these organizations are both concerned about measurement issues and the quality of measurement in educational settings, these collaborations are highly appropriate and fitting.

It is important to note that the editor, the authors, and the chapter reviewers receive no compensation for their hard work. Robert L. Brennan and the various chapter authors and reviewers have worked diligently to assemble an up to date reference book on some of the most critical issues in educational measurement. Many of the topics covered generalize far beyond the confines of education and apply to aspects of measurement in virtually any context. Thus, this work continues the long tradition of *Educational Measurement* in expanding the knowledge base of the entire field. NCME and ACE thank the editor, the authors, and the chapter reviewers for their unselfish contribution.

David Ward, *President*
American Council on Education

James C. Impara, *President*
National Council on Measurement in Education

Editor's Preface

All four editions of *Educational Measurement* have been developed under the aegis of the American Council on Education (ACE), with the last two being collaborative efforts between ACE and the National Council on Measurement in Education (NCME). In the spring of 2002, ACE and NCME began to discuss a revision of *Educational Measurement* under the joint sponsorship of the two organizations. In June 2002, ACE and NCME asked me to assume the role of editor for the fourth edition. Although the contract that formalized the project was not finalized until December 2002, my work on the fourth edition began almost immediately. In particular, I undertook a review of the first three editions of *Educational Measurement,* as well as a review of a substantial body of the literature since the third edition was published, as a basis for constructing a preliminary list of chapters for the fourth edition and potential authors and reviewers.

Also, I identified an Editorial Advisory Committee that was subsequently approved by NCME and ACE. The Editorial Advisory Committee consisted of: Michael A. Baer, who was employed by ACE at that time, Lloyd Bond, Carnegie Foundation, Wendy Bresler, ACE, Linda Crocker, University of Florida, Fritz Drasgow, University of Illinois, Michael T. Kane, National Conference of Bar Examiners, Robert L. Linn, University of Colorado and editor of the third edition, William A. Mehrens, Michigan State University, Cynthia B. Schmeiser, ACT, Inc., and Wendy M. Yen, Educational Testing Service.

The first meeting (and the only formal meeting) of the Editorial Advisory Committee was held in Washington, DC, at the ACE offices on October 29–30, 2002. In addition to the Committee and the editor, other ACE personnel and representatives from the NCME Central Office were in attendance for at least part of the meeting. The Washington meeting was crucial to the project and very productive. The Committee reviewed, discussed, and made suggested revisions to my preliminary list of chapters and chapter lengths. This discussion was followed by an in-depth consideration of potential senior authors for each chapter. Then, potential reviewers for each chapter were identified. I am very grateful for the wise advice and clear support offered to me by the Editorial Advisory Committee.

Shortly after the Washington meeting, I contacted all selected senior authors to ascertain their willingness to write a chapter for the fourth edition. All agreed. Each senior author was given a specified length and told that the writing steps would involve an outline, a draft, and a final submission, with clearly specified deadlines. All authors were given the option of identifying one or more co-authors with my approval. Work progressed steadily but unevenly for the next three years until the final versions of all manuscripts were delivered to me in the fall of 2005.

A comparison of the fourth edition of *Educational Measurement* with the previous three editions illustrates both the enduring nature of many measurement topics and the evolving nature of the field. All editions have chapters devoted to validity, reliability, equating and/or scaling, test development, technology in testing, and measurement in the context of instruction. Also, most editions have chapters dealing with test administration and performance assessment. The titles of some chapters may suggest that a topic is unique to a particular edition, but very often (not always) the issues covered in such chapters are incorporated, at least in part, in other chapters in other editions.

The similarity in chapter titles and/or topics across editions can be quite misleading, however. In particular, although there is overlap in the content coverage of similarly-named chapters, it is almost never the case that a newer chapter is properly viewed as a complete replacement for a previous one. That is as true for the fourth edition as for any other. Many chapters from earlier editions are still relevant in both historical and substantive senses.

The chapters in the fourth edition reflect the authors' perspectives on the considerable changes in the field of educational measurement that have occurred since the third edition was published. Some of these changes occurred as a direct result of developments within the field of educational measurement itself; other changes were substantially influenced by the political and social climate within which the field exists.

This does not mean, however, that there is complete consensus among the authors (or reviewers) of the chapters in the fourth edition, or any previous edition, for that matter. The editor of the fourth edition echoes the warning

by Lindquist (1951) in his preface to the first edition that "the reader of this volume ... should not assume that authorities are fully agreed on all ideas expressed therein" (p. x). As part of my review of the drafts of all chapters, I advised fourth-edition authors about certain inconsistencies across chapters; any differences of opinion that remain reflect professional disagreements that characterize any field of scientific endeavor.

Two typical characteristics of edited books of the size and complexity of the fourth edition are the long time it takes to complete and the large number of persons involved. That has been true of all editions of *Educational Measurement.* The fourth edition took approximately four years to complete, involving direct contributions of one kind or another by over 100 persons.

The Editorial Advisory Committee was particularly helpful to me, as were various NCME and ACE personnel. I particularly thank Linda Crocker, who was President of NCME when the project started, and Susan Slesinger, Executive Editor at Praeger, who advised me throughout the four-year developmental cycle.

Principal credit, of course, is reserved for the authors. In addition, the authors and editor gratefully acknowledge the substantial contributions of numerous reviewers and others who offered comments and otherwise supported the development of the chapters in this volume, as outlined below. Of course, the perspectives and opinions expressed do not necessarily reflect the positions or policies of the authors' employers or funding agencies.

Chapter 1, *Perspectives on the Evolution and Future of Educational Measurement,* was written by the editor and reviewed by Michael T. Kane, National Board of Medical Examiners, and Robert L. Linn, University of Colorado. Additional helpful comments were provided by Michael J. Kolen and Won-Chan Lee, both from the University of Iowa.

Chapter 2, *Validation,* by Michael T. Kane, National Conference of Bar Examiners, was reviewed by Robert L. Linn, University of Colorado, and Pamela A. Moss, University of Michigan. Substantial input was also provided by Janet Kane, independent consultant, and Terence Crooks, University of Otago, Dunedin, NZ.

Chapter 3, *Reliability*, by Edward H. Haertel, Stanford University, was reviewed by Leonard S. Feldt and Won-Chan Lee, both from The University of Iowa.

Chapter 4, *Item Response Theory*, by Wendy M. Yen and Anne R. Fitzpatrick, Educational Testing Service, was reviewed by Mark D. Reckase, Michigan State University, and Peter J. Pashley, Law School Admission Council. Additional helpful comments were provided by Daniel Eignor, Dianne Henderson-Montero, Joanne Lenke, Kevin Meara, Robert Smith, and Matthias von Davier, all from Educational Testing Service, as well as Marc Julian, University of Georgia.

Chapter 5, *Scaling and Norming*, by Michael J. Kolen, The University of Iowa, was reviewed by Eugene Johnson, independent consultant. Additional comments were provided by Ming Mei Wang, Educational Testing Service. Comments on portions of an earlier draft were provided by Zhongmin Cui, NooRee Huh, Seonghoon Kim, Dongmei Li, Kyndra Middleton, Yuki Nozawa, and Ye Tong, all from The University of Iowa.

Chapter 6, *Linking and Equating,* by Paul W. Holland and Neil J. Dorans, Educational Testing Service, was reviewed by Nancy S. Petersen, ACT, Inc., and Mary Pommerich, Defense Manpower Data Center. In addition, helpful comments were provided by Tim Davey, Alina von Davier, Daniel Eignor, Kim Fryer, and Samuel Livingston, all from Educational Testing Service.

Chapter 7, *Test Fairness*, by Gregory Camilli, Rutgers, The State University of New Jersey, was reviewed by Lloyd Bond, Carnegie Foundation, and Bert F. Green, Johns Hopkins University. Additional comments were provided by Susan M. Brookhart, Duquesne University, Wayne J. Camara, College Board, Thomas Van Essen, Educational Testing Service, Leonard S. Feldt, University of Iowa, Cynthia B. Schmeiser, ACT, Inc., Amy E. Schmidt, College Board, Lorrie A. Shepard, University of Colorado, and Kevin G. Welner, University of Colorado.

Chapter 8, *Cognitive Psychology and Educational Assessment,* by Robert J. Mislevy, University of Maryland, was reviewed by David F. Lohman, The University of Iowa, and James W. Pellegrino, University of Illinois at Chicago. Additional comments were provided by John Behrens, Cisco Systems, Inc., Jennifer Cromley, Temple University, Geneva Haertel, SRI International, and Michelle Riconscente, University of Maryland. The chapter builds on work with Russell Almond, Educational Testing Service, and Linda Steinberg on evidence-centered assessment design. The chapter benefited from conversations and projects over the years with colleagues at ETS, CRESST, Cisco Systems, SRI International, the Office of Naval Research, the University of Chicago, the University of Maryland, the Spencer Foundation's Idea of Testing project, and the National Research Council's Committee on the Foundations of Assessment. The work reported here was supported in part by the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

Chapter 9, *Test Development*, by Cynthia B. Schmeiser and Catherine J. Welch, ACT, Inc., was reviewed by Mari Pearlman, Educational Testing Service, and Paul D. Sandifer, independent consultant. Additional helpful comments were provided by Dan Vitale, ACT, Inc.

Chapter 10, *Test Administration, Security, Scoring, and Reporting*, by Allan S. Cohen, University of Georgia, and James A. Wollack, University of Wisconsin, was reviewed by David A. Frisbie, The University of Iowa, and Jeffrey Nellhaus, Massachusetts Department of Education.

Chapter 11, *Performance Assessment,* by Suzanne Lane and Clement A. Stone, University of Pittsburgh, was reviewed by Richard J. Shavelson, Stanford University, Joan Herman, University of California at Los Angeles, and Xiaohong Gao, ACT, Inc. Additional helpful comments were provided by Michael T. Kane, National Conference of Bar Examiners.

Chapter 12, *Setting Performance Standards,* by Ronald K. Hambleton, University of Massachusetts Amherst, and Mary J. Pitoniak, Educational Testing Service, was reviewed

by Susan C. Loomis, National Assessment Governing Board, and Barbara S. Plake, University of Nebraska at Lincoln.

Chapter 13, *Technology and Testing,* by Fritz Drasgow, University of Illinois at Urbana-Champaign, Richard M. Luecht, University of North Carolina at Greensboro, and Randy E. Bennett, Educational Testing Service, was reviewed by Cynthia Parshall, independent consultant, and Stephen G. Sireci, University of Massachusetts Amherst. Additional comments were provided by Judy Spray, ACT, Inc., and James Drasgow, independent consultant.

Chapter 14, *Old, Borrowed and New Thoughts in Second Language Testing,* by Micheline Chalhoub-Deville and Craig Deville, University of North Carolina at Greensboro, was reviewed by Lyle F. Bachman, University of California at Los Angeles, and Richard P. Duran, University of California at Santa Barbara. Additional thanks goes to Caroline Clapham and Nick Saville, University of Cambridge, ESOL Examinations, especially for their input on the history of language testing in the UK.

Chapter 15, *Testing for Accountability in K–12,* by Daniel M. Koretz, Harvard Graduate School of Education, and Laura S. Hamilton, RAND Corporation, was reviewed by Linda Crocker, University of Florida.

Chapter 16, *Standardized Assessment of Individual Achievement in K–12,* by Steve Ferrara and Gerald E. DeMauro, American Institutes for Research, was reviewed by Anthony J. Nitko, University of Arizona, Duncan MacQuarrie, Harcourt Assessment, and Andrew C. Porter, Vanderbilt. Additional support of many types was provided by colleagues at American Institutes for Research.

Chapter 17, *Classroom Assessment,* by Lorrie A. Shepard, University of Colorado, was reviewed by Richard Stiggins, Assessment Training Institute, and Mark R. Wilson, University of California at Berkeley.

Chapter 18, *Higher Education Admissions Testing,* by Rebecca Zwick, University of California, Santa Barbara, was reviewed by Daniel R. Eignor, Educational Testing Service, and E. James Maxey, ACT, Inc. The authors also thank Kathleen O'Neill, Ellen Julian, and Jill Burstein for assistance with sections on the GRE, the MCAT, and the e-Rater program, respectively.

Chapter 19, *Monitoring Educational Progress with Group-Score Assessments,* by John Mazzeo, Stephen Lazer, and Michael J. Zieky, Educational Testing Service, was reviewed by Albert Beaton, Boston College, and Terence Crooks, University of Otago, Dunedin, NZ.

Chapter 20, *Testing for Licensure and Certification in the Professions,* by Brian E. Clauser and Melissa J. Margolis, National Board of Medical Examiners, and Susan M. Case, National Conference of Bar Examiners, was reviewed by Steven M. Downing, University of Illinois at Chicago, and James C. Impara, University of Nebraska at Lincoln. In addition, Polina Harik carefully checked calculations used in the figures and Ronald Nungester, Richard Hawkins, Dave Swanson, and Howard Wainer offered comments on earlier drafts; all are from the National Board of Medical Examiners.

Chapter 21, *Legal and Ethical Issues,* by S. E. Phillips, Consultant, and Wayne J. Camara, College Board, was reviewed by Stan von Mayrhauser, Educational Testing Service, and William A. Mehrens, Michigan State University.

Reference

Lindquist, E. F. (1951). Editor's preface. In E. F. Lindquist (Ed.), *Educational measurement* (pp. vii–xi). Washington, DC: American Council on Education.

Robert L. Brennan

Contents

Part III: Applications

Illustrations

Figures

Tables

1

Perspectives on the Evolution and Future of Educational Measurement

Robert L. Brennan
The University of Iowa

In any scientific field, theorists and practitioners occasionally need to summarize the status of the field, not only to facilitate access to current thinking, but also to enable a reflective consideration of where the field might be headed. That is the principal purpose of this volume. In that spirit, this chapter provides the author's perspectives on various aspects of the evolution and future of educational measurement. These comments are informed by other chapters in this volume, but no claim is made that these comments are always consonant with the perspectives of other authors.

In his introductory chapter to the third edition of *Educational Measurement* Linn (1989a) states:

> A comparison of the current status of educational measurement with that in 1971, when the second edition of this book was published (Thorndike, 1971a), or even with that in 1951, when the first edition appeared (Lindquist, 1951), yields a mixed picture. There are senses in which there has been tremendous change and others in which there has been relatively little. (p. 1)

A similar statement can be made today. While the chapters in this volume reflect the current state-of-the-art in educational measurement, parts of many chapters in the three previous editions are still relevant.

The comments here are organized into two primary sections—measurement theory and measurement practice. The first section traces some of the history of the more technical topics in measurement and concludes with a discussion of the need for more integration of measurement theories. The second section discusses some current issues in measurement practice that are contentious or challenge the field at its boundaries.

1. MEASUREMENT THEORY

To appreciate how measurement theory has evolved since 1950, it is especially instructive to consider how various measurement topics were treated in the three previous editions of *Educational Measurement* (Lindquist, 1951; Linn, 1989b; Thorndike, 1971a) and the five editions of the *Standards for Educational and Psychological Testing* published in 1954, 1966, 1974, 1985, and 1999 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).[1] The *Educational Measurement* and *Standards* citations in the following sections are ordered chronologically—or, nearly so, and followed by comments about how chapters in the current volume represent an evolution of measurement theory.

1.1. Validity

In the first edition of *Educational Measurement* Cureton (1951) states that, "The essential question of test validity is how well a test does the job it is employed to do" (p. 621). He goes on to say:

> Validity has two aspects, which may be termed relevance and reliability. "Relevance" concerns the closeness of agreement between what the test measures and the function that it is used to measure ... Validity is therefore defined in terms of the correlation between the actual test scores and the "true" criterion scores. (pp. 622–623)

Cureton is faithfully reflecting the fact that, "The theory of prediction was very nearly the whole of validity until about 1950" (Cronbach, 1971, p. 443). Afterward, however, everything changed.

1.1.1. The Trinitarian Model

There is no succinct definition of validity in the first edition of the *Standards* (APA, 1954). The document simply states that, "Validity information indicates to the test user the degree to which the test is capable of achieving certain aims" (p. 13). Rather than giving a definition of validity per se, the 1954 *Standards* provides one of the first published discussions of four types of validity: content, predictive, concurrent, and construct.[2] The committee that developed the first edition of the *Standards* in 1954 included both Meehl and Cronbach, with the latter serving as chair. Clearly, they influenced the dramatic shift from prediction as "the whole of validity" to a more nuanced and deeper appreciation of the topic.

In the second edition of the *Standards* (AERA, APA, & NCME, 1966), the perspective on validity is almost identical to that of the 1954 edition, with two exceptions. First, the

validity categories are collapsed from four to three—content, criterion, and construct—which were called "aspects" (p. 12) or "concepts" (p. 14) of validity. (Subsequently, Guion, 1980, called this the trinitarian model of validity.) Second, and much more importantly, the 1966 *Standards* provides an initial discussion of validity of suggested *interpretations*—a notion that resonates to this day.

1.1.2. Inferences and the Centrality of Construct Validity

In the second edition of *Educational Measurement*, Cronbach (1971) states that, "Narrowly considered, *validation* is the process of examining the accuracy of a specific prediction or inference made from a test score ... To explain a test score, one must bring to bear some sort of theory about the causes of the test performance and about its implications" (p. 443). Then, after acknowledging the content/criterion/construct trinitarian model, Cronbach goes on to say, "For purposes of exposition, it is necessary to subdivide *what in the end must be a comprehensive, integrated evaluation of a test*" (p. 445). This statement sounds very much like (nearly) all of validity is construct validity, which Messick (1989, p. 17) later stated, and which is sometimes construed as the "unitary" notion of validity. Subsequently, in the same chapter, Cronbach emphasizes that, "One validates, not a test, but *an interpretation of data arising from a specified procedure*" (p. 447).

The focus on inferences did not escape the committee that authored the third edition of the *Standards* (AERA, APA, & NCME, 1974), which states that, "Questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores or other forms of assessment" (p. 25). Still, the 1974 *Standards* retained discussions of content, criterion, and construct validity—what were called "types" of validity.

Over a decade later, the discussion of validity in the fourth edition of the *Standards* (AERA, APA, & NCME, 1985) is, in its essential features, much like that in the 1974 edition. The fourth edition states, "Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences" (p. 9). There is a slight change in the trinitarian-model terminology; the three categories are called "types of evidence," rather than "types of validity." Perhaps more importantly, however, content-related and criterion-related evidence are viewed as playing a subordinate (or contributing) role to construct-related evidence.

1.1.3. Validity as an Integrated Evaluation

In the third edition of *Educational Measurement* Messick (1989) begins his extensive treatise on validity as follows:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment ... Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use ... Thus the key issues of test validity are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use. (p. 13)

In the context of evolving notions of validity, Messick's treatment of the subject is notable on many levels. Perhaps most importantly, he repeatedly emphasizes that validity is an *integrated evaluative judgment* concerning *inferences* and *social consequences* of test use. He supports this perspective with lengthy discussions that cover numerous fields of inquiry including philosophy of science, in particular. Messick's (1989) chapter, however, is not a treatment of validity that provides much specific guidance to those who would undertake validation studies.

The fifth and most recent edition of the *Standards* (AERA, APA, & NCME, 1999) follows Messick (1989) very closely. It states:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests ... The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. (p. 9)

Subsequent discussion gives considerable attention to notions of construct under-representation and construct irrelevant variance (concepts originally introduced by Cook & Campbell, 1979, and discussed extensively by Messick, 1989), while the content/criterion/construct trinitarian model is essentially replaced by a discussion of sources of validity evidence, including evidence based on consequences of test use.

1.1.4. Validity as Argument

In his treatment of validity in chapter 2 of this volume, Kane provides a perspective on validity that is largely consonant with that of Messick, but there are differences, at least in emphasis. In particular, Kane extends suggestions of Cronbach (1988) and emphasizes a general methodology for validation based on conceptualizing validity as argument. In Kane's formulation, the *validity argument* provides an overall evaluation of the intended interpretations and uses of test scores. The goal is to provide a coherent analysis of all of the evidence for and against proposed interpretations/uses, and to the extent possible, the evidence relevant to plausible competing interpretations. An *interpretative argument* provides a framework for developing a validity argument. In particular, Kane suggests that a convenient basis for discussing the structure of interpretive arguments is to focus on the types of inferences commonly found in test-score interpretations. For example, for a placement testing system, the inferences he discusses are scoring, generalization, extrapolation, and decision making. Many aspects of Kane's perspective on validity are reminiscent of concepts in program evaluation (see, for example, Cronbach, 1982).

Kane's approach to validation can be formulated in terms of the following steps (see Kane, 2001, p. 330):

1. state the proposed interpretation in terms of an interpretative argument, which includes inferences and assumptions involved in the interpretation;
2. create a preliminary version of the validity argument by assembling all available evidence relevant to the plausibility of the interpretive argument;
3. evaluate the most problematic assumptions in detail; and
4. reformulate the interpretive and validity arguments, if necessary, and repeat step three until all inferences in the interpretive argument are considered plausible, or the interpretive argument is rejected.

This is much like the process of theory development in the physical sciences, but Kane's treatment of validity is not influenced solely by philosophy of science. Kane's exposition of validity as argument uses as tools most of the measurement methodologies currently available (particularly generalizability theory) and aims at making validation a more accessible enterprise for educational measurement practitioners. This latter goal is very much consonant with the directive in the 1999 *Standards* that, "the ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user" (p. 112).

1.2. Reliability

Definitions and concepts of validity have evolved considerably since the first edition of *Educational Measurement* was published in 1951. By contrast, the generic definition of reliability has remained largely intact—namely, reliability refers to consistency of scores across replications of a measurement procedure (see Brennan, 2001a). In their treatments of reliability, the various editions of both the *Standards* and *Educational Measurement* differ somewhat in their treatments of types of reliability coefficients (e.g., coefficients of equivalence, stability, and internal consistency), but the primary developments are an increasing attention over time to standard errors of measurement (SEMs) and to generalizability theory (Brennan, 2001b; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

1.2.1. Early 1950s

In the first edition of *Educational Measurement* Thorndike (1951) states: "This tendency toward consistency from one set of measurements to another ... will be designated 'reliability'" (p. 560). Then, almost immediately, Thorndike discussed reliability and analysis of variance (ANOVA). This discussion is too introductory to be characterized as a precursor to generalizability theory, but it might have influenced the editor, who later published an experimental design text (Lindquist, 1953) in which the last chapter foreshadowed the subsequent development of generalizability theory by Cronbach and his colleagues. Interestingly, Cronbach was one of the "collaborators" for Thorndike's chapter, but there is no evidence that the chapter substantially influenced Cronbach's thinking about generalizability theory.

One noteworthy aspect of Thorndike's chapter is what is *not* referenced or even mentioned. For example, there is no reference to Gulliksen's (1950) book, which provided an excellent systematic development of reliability theory up to that time. Also, there is no mention of Cronbach's (1951) paper on coefficient alpha. It seems virtually certain that these publications were overlooked simply because the Thorndike (1951) chapter was largely completed before they were published. (Work on the first edition of *Educational Measurement* began in 1945.)

One undeniable fact about the history of educational measurement is the central role of Cronbach's (1951) paper on coefficient alpha. This paper is widely cited in the social science literature, and probably no statistic related to measurement is reported more frequently. In that sense, the alpha paper has been, and continues to be, extraordinarily influential not only in the field of educational measurement per se but also in many other social science fields. Yet, the emphasis given to Cronbach's alpha is somewhat unfortunate for two reasons. First, as noted by Cronbach (2004) just before his death, coefficient alpha was not particularly novel.[3] Indeed, Cronbach (2004) expressed some embarrassment that his name is uniquely tied to the coefficient.

Second, alpha was never intended by Cronbach to be a universal reliability coefficient. Cronbach had an entirely different goal in mind, as he clearly states in the 1951 paper:

> A ... reason for the symbol is that α is one of six analogous coefficients (to be designated β, γ, δ, etc.) which deal with such other concepts as like-mindedness of persons, stability of scores, etc. (pp. 299–300)

Cronbach abandoned work on "analogous coefficients" when he and his colleagues invented generalizability theory and realized that it was a much richer and more useful approach to conceptualizing and quantifying the influence of different sources of errors on different objects of measurement.

1.2.2. Mid 1950s to Mid 1970s

This two-decade period witnessed the development of the foundations for generalizability theory. In discussing the genesis of the theory, Cronbach (1991) states:

> In 1957 I obtained funds from the National Institute of Mental Health to produce, with Gleser's collaboration, a kind of handbook of measurement theory ... "Since reliability has been studied thoroughly and is now understood," I suggested to the team, "let us devote our first few weeks to outlining that section of the handbook, to get a feel for the undertaking." We learned humility the hard way—the enterprise never got past that topic. Not until 1972 did the book appear ... that exhausted our findings on reliability reinterpreted as generalizability. Even then, we did not exhaust the topic.
>
> When we tried initially to summarize prominent, seemingly transparent, convincingly argued papers on test reliability, the messages conflicted. (pp. 391–392)

To resolve these conflicts, Cronbach and his colleagues devised a rich conceptual framework and married it to analysis of random effects variance components. The net