Part IV
ETS Contributions to Validity

Chapter 16
Research on Validity Theory and Practice at ETS

Michael Kane and Brent Bridgeman

Educational Testing Service (ETS) was founded with a dual mission: to provide high-quality testing programs that would enhance educational decisions and to improve the theory and practice of testing in education through research and development (Bennett 2005; Educational Testing Service 1992). Since its inception in 1947, ETS has consistently evaluated its testing programs to help ensure that they meet high standards of technical and operational quality, and where new theory and new methods were called for, ETS researchers made major contributions to the conceptual frameworks and methodology.

This chapter reviews ETS's contributions to validity theory and practice at various levels of generality, including overarching frameworks (Messick 1988, 1989), more targeted models for issues such as fairness, and particular analytic methodologies (e.g., reliability, equating, differential item functioning). The emphasis will be on contributions to the theory of validity and, secondarily, on the practice of validation rather than on specific methodologies.

16.1 Validity Theory

General conceptions of validity grew out of basic concerns about the accuracy of score meanings and the appropriateness of score uses (Kelley 1927), and they have necessarily evolved over time as test score uses have expanded, as proposed interpretations have been extended and refined, and as the methodology of testing has become more sophisticated.

M. Kane (✉) • B. Bridgeman
Educational Testing Service, Princeton, NJ, USA
e-mail: mkane@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_16

In the first edition of *Educational Measurement* (Lindquist 1951), which was released just after ETS was founded, Cureton began the chapter on validity by suggesting that "the essential question of test validity is how well a test does the job it is employed to do" (Cureton 1951, p. 621) and went on to say that

> validity has two aspects, which may be termed relevance and reliability. … To be valid—that is to serve its purpose adequately—a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to measure. (p. 622)

In the late 1940s and early 1950s, tests tended to be employed to serve two kinds of purposes: providing an indication of the test taker's standing on some attribute (e.g., cognitive ability, personality traits, academic achievement) and predicting future performance in some context.

Given ETS's mission (Bennett, Chap. 1, this volume) and the then current conception of validity (Cureton 1951), it is not surprising that much of the early work on validity at ETS was applied rather than theoretical; it focused on the development of measures of traits thought to be relevant to academic success and on the use of these measures to predict future academic performance. For example, the second Research Bulletin published at ETS (i.e., Frederiksen 1948) focused on the prediction of first-year grades at a particular college.

This kind of applied research designed to support and evaluate particular testing programs continues to be an essential activity at ETS, but over the years, these applied research projects have also generated basic questions about the interpretations of test scores, the statistical methodology used in test development and evaluation, the scaling and equating of scores, the variables to be used in prediction, structural models relating current performance to future outcomes, and appropriate uses of test scores in various contexts and with various populations. In seeking answers to these questions, ETS researchers contributed to the theory and practice of educational measurement by developing general frameworks for validation and related methodological developments that support validation.

As noted earlier, at the time ETS was founded, the available validity models for testing programs emphasized score interpretations in terms of traits and the use of scores as predictors of future outcomes, but over the last seven decades, the concept of validity has expanded. The next section reviews ETS's contributions to the development and validation of trait interpretations, and the following section reviews ETS's contributions to models for the prediction of intended "criterion" outcomes. The fourth describes ETS's contributions to our conceptions and analyses of fairness in testing. The fifth section traces the development of Messick's comprehensive, unified model of construct validity, a particularly important contribution to the theory of validity. The sixth section describes ETS's development of argument-based approaches to validation. A seventh section, on validity research at ETS, focuses on the development of methods for the more effective interpretation and communication of test scores and for the control of extraneous variance. The penultimate section discusses fairness as a core validity concern. The last section provides some concluding comments.

This organization is basically thematic, with each section examining ETS's contributions to the development of aspects of validity theory, but it is also roughly chronological. The strands of the story (trait interpretations, prediction, construct interpretations, models for fairness, Messick's unified model of construct validity, models for the role of consequences of testing, and the development of better methods for encouraging clear interpretations and appropriate uses of test scores) overlap greatly, developed at different rates during different periods, and occasionally folded back on themselves, but there was also a gradual progression from simpler and more intuitive models for validity to more complex and comprehensive models, and the main sections in this chapter reflect this progression.

As noted, most of the early work on validity focused on trait interpretations and the prediction of desired outcomes. The construct validity model was proposed in the mid-1950s (Cronbach and Meehl 1955), but it took a while for this model to catch on. Fairness became a major research focus in the 1970s. In the 1970s and 1980s, Messick developed his unified framework for the construct validity of score interpretations and uses, and the argument-based approaches were developed at the turn of the century.

It might seem appropriate to begin this chapter by defining the term *validity*, but as in any area of inquiry (and perhaps more so than in many other areas of inquiry), the major developments in validity theory have involved changes in what the term means and how it is used. The definition of *validity* has been and continues to be a work in progress. Broadly speaking, validation has always involved an evaluation of the proposed interpretations and uses of test scores (Cronbach 1971; Kane 2006, 2013a; Messick 1989), but both the range of proposed interpretations and the evaluative criteria have gradually expanded.

16.2 Validity of Trait Interpretations

For most of its history from the late nineteenth century to the present, test theory has tended to focus on traits, which were defined in terms of dispositions to behave or perform in certain ways in response to certain kinds of stimuli or tasks, in certain kinds of contexts. Traits were assumed to be personal characteristics with some generality (e.g., over some domain of tasks, contexts, occasions). In the late 1940s and early 1950s, this kind of trait interpretation was being applied to abilities, skills, aptitudes, and various kinds of achievement as well as to psychological traits as such. Trait interpretations provided the framework for test development and, along with predictive inferences, for the interpretation of test scores (Gulliksen 1950a). As theory and methodology developed, *trait interpretations* tended to become more sophisticated in their conceptualizations and in the methods used to estimate the traits. As a result, trait interpretations have come to overlap with *construct interpretations* (which can have more theoretical interpretations), but in this section, we limit ourselves to basic trait interpretations, which involve dispositions to perform in some way in response to tasks of some kind.

Cureton (1951) summarized the theoretical framework for this kind of trait interpretation:

> When the item scores of a set of test-item performances correlate substantially and more or less uniformly with one another, the sum of the item scores (the summary score or test score) has been termed a quasi-measurement. It is a quasi-measurement of "whatever," in the reaction-systems of the individuals, is invoked in common by the test items as presented in the test situation. This "whatever" may be termed a "trait." The existence of the trait is demonstrated by the fact that the item scores possess some considerable degree of homogeneity; that is, they measure in some substantial degree the same thing. We term this "thing" the "trait." (pp. 647–648)

These traits can vary in their content (e.g., achievement in geography vs. anxiety), in their generality (e.g., mechanical aptitude vs. general intelligence), and in the extent to which they are context or population bound, but they share three characteristics (Campbell 1960; Cureton 1951). First, they are basically defined in terms of some relatively specific domain of performance or behavior (with some domains broader than others). Second, the performances or behaviors are assumed to reflect some characteristic of individuals, but the nature of this characteristic is not specified in any detail, and as a result, the interpretation of the trait relies heavily on the domain definition. Third, traits are assumed to be enduring characteristics of individuals, with some more changeable (e.g., achievement in some academic subject) than others (e.g., aptitudes, personality).

Note that the extent to which a trait is enduring is context dependent. Levels of achievement in an academic subject such as geography would be expected to increase while a student is studying the subject and then to remain stable or gradually decline thereafter. A personality trait such as conscientiousness is likely to be more enduring, but even the most stable traits can change over time.

An understanding of the trait (rudimentary as it might be) indicates the kinds of tasks or stimuli that could provide information about it. The test items are designed to reflect the trait, and to the extent possible nothing else, and differences in test scores are assumed to reflect mainly differences in level of the trait.

The general notion of a trait as a (somewhat) enduring characteristic of a person that is reflected in certain kinds of behavior in certain contexts is a basic building block of "folk psychology," and as such, it is ancient (e.g., Solomon was wise, and Caesar was said to be ambitious). As they have been developed to make sense of human behavior over the last century and a half, modern theories of psychology have made extensive use of a wide variety of traits (from introversion to mathematical aptitude) to explain human behavior. As Messick (1989) put it, "a trait is a relatively enduring characteristic of a person—an attribute, process, or disposition—which is consistently manifested to an appropriate degree when relevant, despite considerable variation in the range of settings and circumstances" (p. 15). Modern test theory grew out of efforts to characterize individuals in terms of traits, and essentially all psychometric theories (including classical test theory, generalizability theory, factor analysis, and item response theory) involve the estimation of traits of one kind or another.

From a psychological point of view, the notion of a trait suggests a persistent characteristic of a person that is prior to and independent of any testing program. The trait summarizes (and, in that sense, accounts for) performance or behavior. The trait is not synonymous with any statistical parameter, and it is reasonable to ask whether a parameter estimate based on a particular sample of behavior is an unbiased estimate of the trait of interest. Assuming that the estimate is unbiased, it is also reasonable to ask how precise the estimate is. An assessment of the trait may involve observing a limited range of performances or behaviors in a standardized context and format, but the trait is interpreted in terms of a tendency or disposition to behave in some way or an ability to perform some kinds of tasks in a range of test and non-test contexts. The trait interpretation therefore entails expectations that assessments of the trait using different methods should agree with each other, and assessments of different traits using common methods should not agree too closely (Campbell 1960; Campbell and Fiske 1959).

Traits have two complementary aspects. On one hand, a trait is thought of as an unobservable characteristic of a person, as some latent attribute or combination of such attributes of the person. However, when asked to say what is meant by a trait, the response tends to be in terms of some domain of observable behavior or performance. Thus traits are thought of as unobservable attributes and in terms of typical performance over some domain. Most of the work described in this section focuses on traits as dispositions to behave in certain ways. In a later section, we will focus more on traits as theoretical constructs that are related to domains of behavior or performance but that are defined in terms of their properties as underlying latent attributes or constructs.

16.2.1 ETS's Contributions to Validity Theory for Traits

Trait interpretations of test scores go back at least to the late nineteenth century and therefore predate both the use of the term *validity* and the creation of ETS. However ETS researchers made many contributions to theoretical frameworks and specific methodology for the validation of trait interpretations, including contributions to classical test theory (including reliability theory, standard errors, and confidence intervals), item response theory, equating, factor analysis, scaling, and methods for controlling trait-irrelevant variance. The remainder of this section concentrates on ETS's contributions to the development of these methodologies, all of which seek to control threats to validity.

ETS researchers have been involved in analyzing and measuring a wide variety of traits over ETS's history (Stricker, Chap. 13, this volume), including acquiescence (Messick 1965, 1967), authoritarian attitudes (Messick and Jackson 1958), emotional intelligence (Roberts et al. 2008), cognitive structure (Carroll 1974), response styles (Jackson and Messick 1961; Messick 1991), risk taking (Myers 1965), and social intelligence (Stricker and Rock 1990), as well as various kinds of aptitudes and achievement. ETS researchers have also made major contributions to

the methodology for evaluating the assumptions inherent in trait interpretations and in ruling out factors that might interfere with the intended trait interpretations, particularly in classical test theory (Lord and Novick 1968), theory related to the sampling of target domains (Frederiksen 1984), and item response theory (Lord 1951, 1980).

16.2.2 Classical Test Theory and Reliability

Classical test theory (CTT) is based on trait interpretations, particularly on the notion of a trait score as the expected value over the domain of replications of a measurement procedure. The general notion is that the trait being measured remains invariant over replications of the testing procedure; the test scores may fluctuate to some extent over replications, but the value of the trait is invariant, and fluctuations in observed scores are treated as random errors of measurement. Gulliksen (1950b) used this notion as a starting point for his book, in which he summarized psychometric theory in the late 1940s but used the term *ability* instead of *trait*:

> It is assumed that the gross score has two components. One of these components (T) represents the actual ability of the person, a quantity that will be relatively stable from test to test as long as the tests are measuring the same thing. The other component (E) is an error. (p. 4)

Note that the true scores of CTT are expected values over replications of the testing procedure; they do not refer to an underlying, "real" value of the trait, which has been referred to as a platonic true score to differentiate it from the classical true score. Reliability coefficients were defined in terms of the ratio of true-score variance to observed-score variance, and the precision of the scores was evaluated in terms of the reliability or in terms of standard errors of measurement. Livingston (1972) extended the notion of reliability to cover the dependability of criterion-referenced decisions.

Evidence for the precision of test scores (e.g., standard errors, reliability) supports validity claims in at least three ways. First, some level of precision is necessary for scores to be valid for any interpretation; that is, if the trait estimates have low reliability (i.e., they fluctuate substantially over replications), the only legitimate interpretation of the scores is that they mostly represent error or "noise." Second, the magnitude of the standard error can be considered part of the interpretation of the scores. For example, to say that a test taker has an estimated score of 60 with a standard error of 2 is a much stronger claim than a statement that a test taker has an estimated score of 60 with a standard error of 20. Third, the relationships between the precision of test scores and the number and characteristics of the items in the test can be used to develop tests that are more reliable without sacrificing relevance, thereby improving validity.

Classical test theory was the state of the art in the late 1940s, and as ETS researchers developed and evaluated tests of various traits, they refined old methods and developed new methods within the context of the CTT model (Moses, Chaps. 2 and

3, this volume). The estimation of reliability and standard errors has been an ongoing issue of fundamental importance (Horst 1951; Jöreskog 1971; Keats 1957; Kristof 1962, 1970, 1974; Lord 1955; Novick and Lewis 1967; Tucker 1949). ETS's efforts to identify the implications of various levels of reliability began soon after its inception and have continued since (Angoff 1953; Haberman 2008; Horst 1950a, b; Kristof 1971; Livingston and Lewis 1995; Lord 1956, 1957, 1959).

An important early contribution of ETS researchers to the classical model was the development of conditional standard errors (Keats 1957; Lord 1955, 1956) and of associated confidence intervals around true-score estimates (Gulliksen 1950b; Lord and Novick 1968; Lord and Stocking 1976). Putting a confidence interval around a true-score estimate helps to define and limit the inferences that can be based on the estimate; for example, a decision to assign a test taker to one of two categories can be made without much reservation if a highly conservative confidence interval (e.g., 99%) for a test taker does not include the cutscore between the two categories (Livingston and Lewis 1995). Analyses of the reliability and correlations of subscores can also provide guidance on whether it would be meaningful to report the subscores separately (Haberman 2008).

Evaluations of the precision of test scores serve an important quality-control function, and they can help to ensure an adequate level of precision in the test scores generated by the testing program (Novick and Thayer 1969). Early research established the positive relationship between test length and reliability as well as the corresponding inverse relationship between test length and standard errors (Lord 1956, 1959). That research tradition also yielded methods for maximizing the reliability of composite measures (B.F. Green 1950).

One potentially large source of error in testing programs that employ multiple forms of a test (e.g., to promote security) is variability in content and statistical characteristics (particularly test difficulty) across different forms of the test, involving different samples of test items. Assuming that the scores from the different forms are to be interpreted and used interchangeably, it is clearly desirable that each test taker's score be more or less invariant across the forms, but this ideal is not likely to be met exactly, even if the forms are developed from the same specifications. Statistical equating methods are designed to minimize the impact of form differences by adjusting for differences in operating characteristics across the forms. ETS researchers have made major contributions to the theory and practice of equating (Angoff 1971; Holland 2007; Holland and Dorans 2006; Holland and Rubin 1982; Lord and Wingersky 1984; Petersen 2007; Petersen et al. 1989; A.A. von Davier 2011; A.A. von Davier et al. 2004). In the absence of equating, form-to-form differences can introduce substantial errors, and equating procedures can reduce this source of error.

On a more general level, ETS researchers have played major roles in developing the CTT model and in putting it on firm foundations (Lord 1965; Novick 1965). In 1968, Frederick Lord and Melvin Novick formalized and summarized most of what was known about the CTT model in their landmark book *Statistical Theories of Mental Test Scores*. They provided a very sophisticated statement of the classical test-theory model and extended it in many directions.

16.2.3 Adequate Sampling of the Trait

Adequate sampling of the trait domain requires a clear definition of the domain, and ETS researchers have devoted a lot of attention to developing a clear understanding of various traits and of the kinds of performances associated with these traits (Ebel 1962). For example, Dwyer et al. (2003) defined *quantitative reasoning* as "the ability to analyze quantitative information" (p. 13) and specified that its domain would be restricted to quantitative tasks that would be new to the student (i.e., would not require methods that the test takers had been taught). They suggested that quantitative reasoning includes six more specific capabilities: (a) understanding quantitative information presented in various formats, (b) interpreting and drawing inferences from quantitative information, (c) solving novel quantitative problems, (d) checking the reasonableness of the results, (e) communicating quantitative information, and (f) recognizing the limitations of quantitative methods. The quantitative reasoning trait interpretation assumes that the tasks do not require specific knowledge that is not familiar to all test takers and, therefore, any impact that such knowledge has on the scores would be considered irrelevant variance.

As noted earlier, ETS has devoted a lot of attention to developing assessments that reflect traits of interest as fully as possible (Lawrence and Shea 2011). Much of this effort has been devoted to more adequately sampling the domains associated with the trait, and thereby reducing the differences between the test content and format and the broader domain associated with the trait (Bejar and Braun 1999; Frederiksen 1984). For example, the "in basket" test (Frederiksen et al. 1957) was designed to evaluate how well managers could handle realistic versions of management tasks that required decision making, prioritizing, and delegating. Frederiksen (1959) also developed a test of creativity in which test takers were presented with descriptions of certain results and were asked to list as many hypotheses as they could to explain the results. Frederiksen had coauthored the chapter on performance assessment in the first edition of *Educational Measurement* (Ryans and Frederiksen 1951) and consistently argued for the importance of focusing assessment on the kinds of performance that are of ultimate interest, particularly in a landmark article, "The Real Test Bias: Influences of Testing on Teaching and Learning" (Frederiksen 1984). More recently, ETS researchers have been developing a performance-based program of Cognitively Based Assessment *of*, *for*, and *as* Learning (the *CBAL*® initiative) that elicits extended performances (Bennett 2010; Bennett and Gitomer 2009). For CBAL, and more generally for educational assessments, positive changes in the traits are the goals of instruction and assessment, and therefore the traits being assessed are not expected to remain the same over extended periods.

The evidence-centered design (ECD) approach to test development, which is discussed more fully later, is intended to promote adequate sampling of the trait (or construct) by defining the trait well enough up front to get a good understanding of the kinds of behaviors or performance that would provide the evidence needed to draw conclusions about the trait (Mislevy et al. 1999, 2002). To the extent that the

testing program is carefully designed to reflect the trait of interest, it is more likely that the observed behaviors or performances will adequately achieve that end.

Based on early work by Lord (1961) on the estimation of norms by item sampling, matrix sampling approaches, in which different sets of test tasks are taken by different subsamples of test takers, have been developed to enhance the representativeness of the sampled test performances for the trait of interest (Mazzeo et al. 2006; Messick et al. 1983). Instead of drawing a single sample of tasks that are administered to all test takers, multiple samples of tasks are administered to different subsamples of test takers. This approach allows for a more extensive sampling of content in a given amount of testing time. In addition, because it loosens the time constraints on testing, the matrix sampling approach allows for the use of a wider range of test tasks, including performance tasks that require substantial time to complete. These matrix sampling designs have proven to be especially useful in large-scale monitoring programs like the National Assessment of Educational Progress (NAEP) and in various international testing programs (Beaton and Barone, Chap. 8, Kirsch et al. Chap. 9, this volume).

16.2.4 Factor Analysis

Although a test may be designed to reflect a particular trait, it is generally the case that the test scores will be influenced by many characteristics of the individuals taking the test (e.g., motivation, susceptibility to distractions, reading ability). To the extent that it is possible to control the impact of test-taker characteristics that are irrelevant to the trait of interest, it may be possible to interpret the assessment scores as relatively pure measures of that focal trait (French 1951a, b, 1954, 1963). More commonly, the assessment scores may also intentionally reflect a number of test-taker characteristics that, together, compose the trait. That is, broadly defined traits that are of practical interest may involve a number of more narrowly defined traits or factors that contribute to the test taker's performance. For example, as noted earlier, Dwyer et al. (2003) defined the performance domain for quantitative reasoning in terms of six capabilities, including understanding quantitative information, interpreting quantitative information, solving quantitative problems, and estimating and checking answers for reasonableness. In addition, most trait measures require ancillary abilities (e.g., the ability to read) that are needed for effective performance in the assessment context.

In interpreting test scores, it is generally helpful to develop an understanding of how different characteristics are related to each other. Factor analysis models have been widely used to quantify the contributions of different underlying characteristics, or "factors," to assessment scores, and ETS researchers have played a major role in the development of various factor-analytic methods (Moses, Chaps. 2 and 3, this volume), in part because of their interest in developing a variety of cognitive and noncognitive measures (French 1951a, b, 1954).

Basic versions of exploratory factor analysis were in general use when ETS was formed, but ETS researchers contributed to the development and refinement of more

sophisticated versions of these methods (Browne 1968; B.F. Green 1952; Harman 1967; Lord and Novick 1968; Tucker 1955). Exploratory factor analysis makes it possible to represent the relationships (e.g., correlations or covariances) among observed scores on a set of assessments in terms of a statistical model describing the relationships among a relatively small number of underlying dimensions, or factors. The factor models decompose the observed total scores on the tests into a linear combination of factor scores, and they provide quantitative estimates of the relative importance of the different factors in terms of the variance explained by the factor.

By focusing on the traits as latent dimensions or factors or as some composite of more basic latent factors, and by embedding these factors within a web of statistical relationships, exploratory factor analysis provided a rudimentary version of the kind of nomological networks envisioned by Cronbach and Meehl (1955). The utility of exploratory analyses for explicating appropriate interpretations of test scores was enhanced by an extended research program at ETS to develop sets of reference measures that focused on particular basic factors (Ekstrom et al. 1979; French 1954; French et al. 1963). By including the reference tests with a more broadly defined trait measure, it would be possible to evaluate the factor structure of the broadly defined trait in terms of the reference factors.

As in other areas of theory development, the work done on factor analysis by ETS researchers tended to grow out of and be motivated by concerns about the need to build assessments that reflected certain traits and to evaluate how well the assessment actually reflected those traits. As a result, ETS's research on exploratory factor analysis has involved a very fruitful combination of applied empirical studies of score interpretations and sophisticated theoretical modeling (Browne 1968; French 1951a, b; Harman 1967; Lord and Novick 1968).

A major contribution to the theory and practice of validation that came out of research at ETS is confirmatory factor analysis (Jöreskog 1967, 1969; Jöreskog and Lawley 1967; Jöreskog and van Thillo 1972). As its name indicates, exploratory factor analysis does not propose strong constraints a priori; the analysis essentially partitions the observed-score variances by using statistical criteria to fit the model to the data. In a typical exploratory factor analysis, theorizing tends to occur after the analysis, as the resulting factor structure is used to suggest plausible interpretations for the factors. If reference factors are included in the analysis, they can help orient the interpretation.

In confirmatory factor analysis (CFA), a factor model is specified in advance by putting constraints on the factor structure, and the constrained model is fit to the data. The constraints imposed on the model are typically based on a priori theoretical assumptions, and the empirical data are used to check the hypotheses built into the models. As a result, CFAs can provide support for theory-based hypotheses or can result in refutations of some or all or the theoretical conjectures (Jöreskog 1969). This CFA model was extended as the basis for structural equation modeling (Jöreskog and van Thillo 1972). To the extent that the constraints incorporate theoretical assumptions, CFAs go beyond simple trait interpretations into theory-based construct interpretations.

CFA is very close in spirit and form to the nomological networks of Cronbach and Meehl (1955). In both cases, there are networks of hypothesized relationships between constructs (or latent variables), which are explicitly defined a priori and which may be extensive, and there are proposed measures of at least some of the constructs. Given specification of the network as a confirmatory factor model (and adequate data), the hypotheses inherent in the network can be checked by evaluating the fit of the model to the data. If the model fits, the substantive assumptions (about relationships between the constructs) in the model and the validity of the proposed measures of the constructs are both supported. If the model does not fit the data, either the substantive assumptions and/or the validity of the measures is likely to be questioned. As is the case in the classic formulation of the construct validity model (Cronbach and Meehl 1955), the substantive theory and the assessments are initially validated (or invalidated) holistically as a network of interrelated assumptions. If the constrained model fails to fit the data, the data can be examined to identify potential weaknesses in the network. In addition, the model fit can be compared to the fit of alternate models that make different (perhaps stronger or weaker) assumptions.

16.2.5 Latent Traits

Two major developments in test theory in the second half of the twentieth century (the construct validity model and latent trait theory) grew out of attempts to make the relationship between observed behaviors or performances and the relevant traits more explicit, and ETS researchers played major roles in both of these developments (see Carlson and von Davier, Chap. 5, this volume). Messick (1975, 1988, 1989) elaborated the construct validity model of Cronbach and Meehl (1955), which sought to explicate the relationships between traits and observed assessment performances through substantive theories that would relate trait scores to the constructs in a theory and to other trait scores attached to the theory. Item response theory (IRT) deployed measurement models to specify the relationships between test performances and postulated latent traits and to provide statistical estimates of these traits (Lord 1951). Messick's contributions to construct validity theory will be discussed in detail later in this chapter. In this section, we examine contributions to IRT and the implications of these developments for validity.

In their seminal work on test theory, Lord and Novick (1968) used trait language to distinguish true scores from errors:

> Let us suppose that we repeatedly administer a given test to a subject and thus obtain a measurement each day for a number of days. Further, let us assume that with respect to the particular *trait* the test is designed to measure, the person does not change from day to day and that successive measurements are unaffected by previous measurements. Changes in the environment or the *state* of the person typically result in some day-to-day variation in the measurements which are obtained. We may view this variation as the result of errors of measurement of the underlying trait characterizing the individual, or we may view it as a representation of a real change in this trait. (pp. 27–28)

In models for true scores, the true score captures the enduring component in the scores over repeated, independent testing, and the "random" fluctuations around this true score are relegated to error.

Lord and Novick (1968) also used the basic notion of a trait to introduce latent traits and item characteristic functions:

> Any theory of latent traits supposes that an individual's behavior can be accounted for, to a substantial degree, by defining certain human characteristics called *traits*, quantitatively estimating the individual's standing on each of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations. (p. 358)

Within the context of the statistical model, the latent trait accounts for the test performances, real and possible, in conjunction with item or task parameters. The latent trait has model-specific meaning and a model-specific use; it captures the enduring contribution of the test taker's "ability" to the probability of success over repeated, independent performances on different tasks.

Latent trait models have provided a richer and in some ways firmer foundation for trait interpretations than offered by classical test theory. One motivation for the development of latent trait models (Lord 1951) was the realization that number-right scores and simple transformations of such scores would not generally yield the defining property of traits (i.e., invariance over measurement operations). The requirement that task performance data fit the model can also lead to a sharpening of the domain definition, and latent trait models can be helpful in controlling random errors by facilitating the development of test forms with optimal statistical properties and the equating of scores across different forms of a test.

A model-based trait interpretation depends on empirical evidence that the statistical model fits the data well enough; if it does, we can have confidence that the test scores reflect the trait conceived of as "whatever … is invoked in common by the test items" (Cureton 1951, p. 647). The application of a CTT or latent trait model to student responses to generate estimates of a true score or a latent trait does not in itself justify the interpretation of scores in terms of a construct that causes and explains the task performances, and it does not necessarily justify inferences to any nontest performance. A stronger interpretation in terms of a psychological trait that has implications beyond test scores requires additional evidence (Messick 1988, 1989). We turn to such construct interpretations later in this chapter.

16.2.6 Controlling Irrelevant Variance

As is the case in many areas of inquiry, a kind of negative reasoning can play an important role in validation of trait interpretations. Tests are generally developed to yield a particular score interpretation and often a particular use, and the test development efforts make a case for the interpretation and use (Mislevy et al. 2002). Once this initial positive case has been made, it can be evaluated by subjecting it to

empirical challenge. We can have confidence in claims that have survived all serious challenges.

To the extent that an alternate proposal is as plausible, or more plausible, than a proposed trait interpretation, we cannot have much confidence in the intended interpretation. This notion, which is a fundamental methodological precept in science (Popper 1965), underlies, for example, multitrait–multimethod analyses (D. T. Campbell and Fiske 1959) and the assumption that reliability is a necessary condition for validity. As a result, to the extent that we can eliminate alternative interpretations of test scores, the proposed interpretation becomes more plausible, and if we can eliminate all plausible rivals for a proposed trait interpretation, we can accept that interpretation (at least for the time being).

In most assessment contexts, the question is not whether an assessment measures the trait or some alternate variable but rather the extent to which the assessment measures the trait of interest and is not overly influenced by sources of irrelevant variance. In their efforts to develop measures of various traits, ETS researchers have examined many potential sources of irrelevant variance, including anxiety (French 1962; Powers 1988, 2001), response styles (Damarin and Messick 1965), coaching (Messick 1981b, 1982a; Messick and Jungeblut 1981), and stereotype threat (Stricker 2008; Stricker and Bejar 2004; Stricker and Ward 2004). Messick (1975, 1989) made the evaluation of plausible sources of irrelevant variance a cornerstone of validation, and he made the evaluation of construct-irrelevant variance and construct underrepresentation central concerns in his unified model of validity.

It is, of course, desirable to neutralize potential sources of irrelevant variance before tests are administered operationally, and ETS has paid a lot of attention to the development and implementation of item analysis methodology, classical and IRT-based, designed to minimize irrelevant variance associated with systematic errors and random errors. ETS has played a particularly important role in the development of methods for the detection of differential item functioning (DIF), in which particular items operate inconsistently across groups of test takers while controlling for ability and thereby introduce systematic differences that may not reflect real differences in the trait of interest (Dorans, Chap. 7, this volume, 1989, 2004; Dorans and Holland 1993; Holland and Wainer 1993; Zieky 1993, 2011).

Trait interpretations continue to play a major role in the interpretation and validation of test scores (Mislevy 2009). As discussed earlier, trait interpretations are closely tied to domains of possible test performances, and these domains provide guidance for the development of assessment procedures that are likely to support their intended function. In addition, trait interpretations can be combined with substantive assumptions about the trait and the trait's relationships to other variables, thus going beyond the basic trait interpretation in terms of a domain of behaviors or performances to an interpretation of a theoretical construct (Messick 1989; Mislevy et al. 2002).

16.3 Validity of Score-Based Predictions

Between 1920 and 1950, test scores came to be used to predict future outcomes and to estimate concurrent criteria that were of practical interest but were not easily observed, and the validity of such criterion-based interpretations came to be evaluated mainly in terms of how well the test scores predicted the criterion (Angoff 1988; Cronbach 1971; Kane 2012; Messick 1988, 1989; Zwick 2006). In the first edition of *Educational Measurement*, which was written as ETS was being founded, Cureton (1951) associated validity with "the correlation between the actual test scores and the 'true' criterion score" (p. 623), which would be estimated by the correlation between the test scores and the criterion scores, with an adjustment for unreliability in the criterion.

The criterion variable of interest was assumed to have a definite value for each person, which was reflected by the criterion measure, and the test scores were to "predict" these values as accurately as possible (Gulliksen 1950b). Given this interpretation of the test scores as stand-ins for the true criterion measure, it was natural to evaluate validity in terms of the correlation between test scores and criterion scores:

> Reliability has been regarded as the correlation of a given test with a parallel form. Correspondingly, the validity of a test is the correlation of the test with some criterion. In this sense a test has a great many different "validities." (Gulliksen 1950b, p. 88)

The criterion scores might be obtained at about the same time as the test scores ("concurrent validity"), or they might be a measure of future performance (e.g., on the job, in college), which was not available at the time of testing ("predictive validity"). If a good criterion were available, the criterion model could provide simple and elegant estimates of the extent to which scores could be used to estimate or predict criterion scores (Cureton 1951; Gulliksen 1950b; Lord and Novick 1968). For admissions, placement, and employment, the criterion model is still an essential source of validity evidence. In these applications, criterion-related inferences are core elements in the proposed interpretations and uses of the test scores. Once the criterion is specified and appropriate data are collected, a criterion-based validity coefficient can be estimated in a straightforward way.

As noted earlier, the criterion model was well developed and widely deployed by the late 1940s, when ETS was founded (Gulliksen 1950b). Work at ETS contributed to the further development of these models in two important ways: by improving the accuracy and generality of the statistical models and frameworks used to estimate various criteria (N. Burton and Wang 2005; Moses, Chaps. 2 and 3, this volume) and by embedding the criterion model in a more comprehensive analysis of the plausibility of the proposed interpretation and use of test scores (Messick 1981a, 1989). The criterion model can be implemented more or less mechanically once the criterion has been defined, but the specification of the criterion typically involves value judgments and a consideration of consequences (Messick 1989).

Much of the early research at ETS addressed the practical issues of developing testing programs and criterion-related validity evidence, but from the beginning,

researchers were also tackling more general questions about the effective use of standardized tests in education. The criterion of interest was viewed as a measure of a trait, and the test was conceived of as a measure of another trait that was related to the criterion trait, as an aptitude is related to subsequent achievement. As discussed more fully in a later section, ETS researchers conducted extensive research on the factors that tend to have an impact on the correlations of predictors (particularly *SAT*® scores) with criteria (e.g., first-year college grades), which served as measures of academic achievement (Willingham et al. 1990).

In the 1940s and 1950s, there was a strong interest in measuring both cognitive and noncognitive traits (French 1948). One major outcome of this extensive research program was the finding that cognitive measures (test scores, grades) provided fairly accurate predictions of performance in institutions of higher education and that the wide range of noncognitive measures that were evaluated did not add much to the accuracy of the predictions (Willingham et al. 1990).

As noted by Zwick (2006), the validity of tests for selection has been judged largely in terms of how well the test scores can predict some later criterion of interest. This made sense in 1950, and it continues to make sense into the twenty-first century. The basic role of criterion-related validity evidence in evaluating the accuracy of such predictions continues to be important for the validity of any interpretation or use that relies on predictions of future performance (Kane 2013a), but these paradigm cases of prediction now tend to be evaluated in a broader theoretical context (Messick 1989) and from a broader set of perspectives (Dorans 2012; Holland 1994; Kane 2013b). In this broader context, the accuracy of predictions continues to be important, but concerns about fairness and utility are getting more attention than they got before the 1970s.

16.4 Validity and Fairness

Before the 1950s, the fairness of testing programs tended to be evaluated mainly in terms of equivalent or comparable treatment of test takers. This kind of procedural fairness was supported by standardizing test administration, materials, scoring, and conditions of observation, as a way of eliminating favoritism or bias; this approach is illustrated in the civil service testing programs, in licensure programs, and in standardized educational tests (Porter 2003). It is also the standard definition of fairness in sporting events and other competitions and is often discussed in terms of candidates competing on "a level playing field." Before the 1950s, this very basic notion of fairness in testing programs was evaluated mainly at the individual level; each test taker was to be treated in the same way, or if some adjustment were necessary (e.g., due to a candidate's disability or a logistical issue), as consistently as possible. In the 1950s, 1960s, and 1970s, the civil rights movement, legislation, and litigation raised a broader set of fairness issues, particularly issues of fair treatment of groups that had suffered discrimination in the past (Cole and Moss 1989; Willingham 1999).

With respect to the treatment of groups, concerns about fairness and equal opportunity prior to this period did exist but were far more narrowly defined. One of the goals of James Conant and others in promoting the use of the Scholastic Aptitude Test was to expand the pool of students admitted to major universities by giving all high school students an opportunity to be evaluated in terms of their aptitude and not just in terms of the schools they attended or the curriculum that they had experienced. As president of Harvard in the 1930s, Conant found that most of Harvard's students were drawn from a small set of elite prep schools and that the College Board examinations, as they then existed, evaluated mastery of prep school curricula (Bennett, Chap. 1, this volume): "For Conant, Harvard admission was being based largely on ability to pay. If a student could not afford to attend prep school, that student was not going to do well on the College Boards, and wasn't coming to Harvard" (p. 5). In 1947, when ETS was founded, standardized tests were seen as a potentially important tool for improving fairness in college admissions and other contexts, at least for students from diverse economic backgrounds. The broader issues of adverse impact and fairness as they related to members of ethnic, racial, and gender groups had not yet come into focus.

Those broader issues of racial, ethnic, and gender fairness and bias moved to center stage in the 1960s:

> Hard as it now may be to imagine, measurement specialists more or less discovered group-based test fairness as a major issue only some 30 years ago. Certainly, prior to that time, there was discussion of the cultural fairness of a test and its appropriateness for some examinees, but it was the Civil Rights movement in the 1960s that gave social identity and political dimension to the topic. That was the period when hard questions were first asked as to whether the egalitarian belief in testing was justified in the face of observed subgroup differences in test performance. The public and test specialists alike asked whether tests were inherently biased against some groups, particularly Black and Hispanic examinees. (Willingham 1999, p. 214)

As our conceptions of fairness and bias in testing expanded between the 1960s and the present, ETS played a major role in defining the broader notions of fairness and bias in testing. ETS researchers developed frameworks for evaluating fairness issues, and they developed and implemented methodology to control bias and promote fairness. These frameworks recognized the value of consistent treatment of individual test takers, but they focused on a more general conception of equitable treatment of individuals and groups (J. Campbell 1964; Anne Cleary 1968; Cleary and Hilton 1966; Cole 1973; Cole and Moss 1989; Dorans, Chap. 7, this volume; Frederiksen 1984; Linn 1973, 1975, 1976; Linn and Werts 1971; Messick 1975, 1980, 1989; Wild and Dwyer 1980; Willingham and Cole 1997; Xi 2010).

16.4.1 Fairness and Bias

Although the terms *fairness* and *bias* can be interpreted as covering roughly the same ground, with *fairness* being defined as the absence of bias, fairness often reflects a broader set of issues, including the larger issues of social equity. In contrast, *bias* may be given a narrower and more technical interpretation in terms of irrelevant factors that distort the interpretation of test scores:

> The word fairness suggests fairness that comes from impartiality, lacking in prejudice or favoritism. This implies that a fair test is comparable from person to person and group to group. Comparable in what respect? The most reasonable answer is validity, since validity is the raison d'etre of the entire assessment enterprise. (Willingham 1999, p. 220)

In its broadest uses, fairness tends to be viewed as an ethical and social issue concerned with "the justice and impartiality inherent in actions" (Willingham 1999, p. 221). Bias, conversely, is often employed as a technical concept, akin to the notion of bias in the estimation of a statistical parameter. For example, Cole and Moss (1989) defined bias as the "differential validity of a particular interpretation of a test score for any definable, relevant group of test takers" (p. 205).

Standardized testing programs are designed to treat all test takers in the same way (or if accommodations are needed, in comparable ways), thereby eliminating as many sources of irrelevant variance as possible. By definition, to the extent that testing materials or conditions are not standardized, they can vary from test taker to test taker and from one test administration to another, thereby introducing irrelevant variance, or bias, into test scores. Much of this irrelevant variance would be essentially random, but some of it would be systematic in the sense that some test scores (e.g., those from a test site with an especially lenient or especially severe proctor) would be consistently too high or too low. Standardization also tends to control some kinds of intentional favoritism or negative bias by mandating consistent treatment of all test takers. Test scores that consistently underestimate or overestimate the variable of interest for a subgroup for any reason are said to be biased, and standardization tends to control this kind of bias, whether it is inadvertent or intentional.

ETS and other testing organizations have developed systematic procedures designed to identify and eliminate any aspects of item content or presentation that might have an undue effect on the performance of some test takers: "According to the guidelines used at ETS, for example, the test 'must not contain language, symbols, words, phrases, or examples that are generally regarded as sexist, racist, or otherwise potentially offensive, inappropriate, or negative toward any group'" (Zwick 2006, p. 656). Nevertheless, over time, there was a growing realization that treating everyone in the same way does not necessarily ensure fairness or lack of bias. It is a good place to start (particularly as a way to control opportunities for favoritism, racism, and other forms of more or less overt bias), but it does not fully resolve the issue. As Turnbull (1951) and others recognized from mid-century, fairness depends on the appropriateness of the uses of test scores, and test scores that provide unbiased measures of a particular set of skills may not provide unbiased measures of a broader domain of skills needed in some context (e.g., in an occupation

or in an educational program). In such cases, those test scores may not provide a fair basis for making decisions about test takers (Shimberg 1981, 1982, 1990).

Over the last 65 years or so, ETS researchers have been active in investigating questions about bias and fairness in testing, in defining issues of fairness and bias, and in developing approaches for minimizing bias and for enhancing fairness. Many of the issues are still not fully resolved, in part because questions of bias depend on the intended interpretation and because questions of fairness depend on values.

16.4.2 Adverse Impact and Differential Prediction

Unless we are willing to assume, a priori, that there are no differences between groups in the characteristic being measured, simple differences between groups in average scores or the percentages of test takers achieving some criterion score do not necessarily say anything about the fairness of test scores or of score uses. In 1971, the U.S. Supreme Court, in *Griggs v. Duke Power Co.*, struck down some employment practices at the Duke Power Company that had led to substantially different hiring rates between Black and White applicants, and in its decision, the Court relied on two concepts, adverse impact and business necessity, that have come to play an important role in discussions of possible bias in score-based selection programs. *Adverse impact* occurs if a protected group (defined by race, ethnicity, or gender, as specified in civil rights legislation) has a substantially lower rate of selection, certification, or promotion compared to the group with the highest rate. A testing program has *business necessity* if the scores are shown to be related to some important outcome (e.g., some measure of performance on the job). A testing program with adverse impact against one or more protected groups was required to demonstrate business necessity for the testing program; if there was no adverse impact, there was no requirement to establish business necessity. Employers and other organizations using test scores for selection would either have to develop selection programs that had little adverse impact or would have to demonstrate business necessity (Linn 1972). In *Griggs*, Duke Power's testing program was struck down because it had substantial adverse impact, and the company had made no attempt to investigate the relationship between test scores and performance on the job.

Although the terminology of *adverse impact* and *business necessity* was not in common use before *Griggs*, the notion that test scores can be considered fair if they reflect real differences in performance, even if they also suffer from adverse impact, was not new. Turnbull (1951) had pointed out the importance of evaluating fairness in terms of the proposed interpretation and use of the scores:

> That method is to define the criterion to which a test is intended to relate, and then to justify inter-group equality or inequality of test scores on the basis of its effect on prediction. It is necessarily true that an equality of test scores that would signify fairness of measurement for one criterion on which cultural groups performed alike would signify unfairness for another criterion on which group performance differed. Fairness, like its amoral brother, validity, resides not in tests or test scores but in the relation of test scores to criteria. (pp. 148–149)

Adverse impact does not necessarily say much about fairness, but it does act as a trigger that suggests that the relationship between test scores and appropriate criteria be evaluated (Dorans, Chap. 7, this volume; Linn 1973, 1975; Linn and Werts 1971; Messick 1989).

By 1971, when the Griggs decision was rendered, Cleary (1968) had already published her classic study of differential prediction, which was followed by a number of differential-prediction studies at ETS and elsewhere. The Cleary model stipulated that

> a test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. (p. 115)

The Cleary criterion is simple, clear, and direct; if the scores underpredict or overpredict the relevant criterion, the predictions can be considered biased. Note that although Cleary talked about the test being biased, her criterion applies to the predictions based on the scores and not directly to the test or test scores. In fact, the predictions can be biased in Cleary's sense without having bias in the test scores, and the predictions can be unbiased in Cleary's sense while having bias in the test scores (Zwick 2006). Nevertheless, assuming that the criterion measure is appropriate and unbiased (which can be a contentious assumption in many contexts; e.g., see Linn 1976; Wild and Dwyer 1980), the comparison of regressions made perfect sense as a way to evaluate predictive bias.

However, as a criterion for evaluating bias in the test scores, the comparison of regression lines is problematic for a number of reasons. Linn and Werts (1971) pointed out two basic statistical problems with the Cleary model; the comparisons of the regression lines can be severely distorted by errors of measurement in the independent variable (or variables) and by the omission of relevant predictor variables. Earlier, Lord (1967) had pointed to an ambiguity in the interpretation of differential-prediction analyses for groups with different means on the two measures, if the measures had less than perfect reliability or relevant predictors had been omitted.

In the 1970s, a concerted effort was made by many researchers to develop models of fairness that would make it possible to identify and remove (or at least ameliorate) group inequities in score-based decision procedures, and ETS researchers were heavily involved in these efforts (Linn 1973, 1984; Linn and Werts 1971; Myers 1975; Petersen and Novick 1976). These efforts raised substantive questions about what we might mean by fairness in selection, but by the early 1980s, interest in this line of research had declined for several reasons.

First, a major impetus for the development of these models was the belief in the late 1960s that at least part of the explanation for the observed disparities in test scores across groups was to be found in the properties of the test. The assumption was that cultural differences and differences in educational and social opportunities caused minority test takers to be less familiar with certain content and to be less adept at taking objective tests, and therefore the test scores were expected to under-

predict performance in nontest settings (e.g., on the job, in various educational programs). Many of the fairness models were designed to adjust for inequities (defined in various ways) that were expected to result from the anticipated underprediction of performance. However, empirical results indicated that the test scores did not underpredict the scores of minority test takers, but rather overpredicted the performance of Black and Hispanic students on standard criteria, particularly first-year grade point average (GPA) in college (Cleary 1968; Young 2004; Zwick 2006). The test scores did underpredict the scores of women, but this difference was due in part to differences in courses taken (Wild and Dwyer 1980; Zwick 2006).

Second, Petersen and Novick (1976) pointed out some basic inconsistencies in the structures of the fairness models and suggested that it was necessary to explicitly incorporate assumptions about relative utilities of different outcomes for different test takers to resolve these discrepancies. However, it was not clear how to specify such utilities, and it was especially not clear how to get all interested stakeholders to agree on a specific set of such utilities.

As the technical difficulties mounted (Linn 1984; Linn and Werts 1971; Petersen and Novick 1976) and the original impetus for the development of the models (i.e., underprediction for minorities) turned out to be wrong (Cleary 1968; Linn 1984), interest in the models proposed to correct for underprediction faded.

An underlying concern in evaluating fairness was (and is) the acknowledged weaknesses in the criterion measures (Wild and Dwyer 1980). In addition to being less reliable than the tests being evaluated, and in representing proxy measures of success that are appealing in large part because of their ready availability, there is evidence that the criteria are, themselves, not free of bias (Wild and Dwyer 1980).

One major result of this extended research program is a clear realization that fairness and bias are very complex, multifaceted issues that cannot be easily reduced to a formal model of fairness or evaluated by straightforward statistical analyses (Cole and Moss 1989; Messick 1989; Wild and Dwyer 1980): "The institutions and professionals who sponsor and use tests have one view as to what is fair; examinees have another. They will not necessarily always agree, though both have a legitimate claim" (Willingham 1999, p. 224). Holland (1994) and Dorans (2012) suggested that analyses of test score fairness should go beyond the measurement perspective, which tends to focus on the elimination or reduction of construct-irrelevant variance (or measurement bias), to include the test taker's perspective, which tends to view tests as "contests," and Kane (2013b) has suggested adding an institutional perspective, which has a strong interest in eliminating any identifiable source of bias but also has an interest in reducing adverse impact, whether it is due to an identifiable source of bias or not.

16.4.3 Differential Item Functioning

ETS played a major role in the introduction of DIF methods as a way to promote fairness in testing programs (Dorans and Holland 1993; Holland and Thayer 1988). These methods identify test items that, after matching on an estimate of the attribute

of interest, are differentially difficult or easy for a target group of test takers, as compared to some reference group. ETS pioneered the development of DIF methodology, including the development of the most widely used methods, as well as investigations of the statistical properties of these methods, matching variables, and sample sizes (Dorans, Chap. 7, this volume; Holland and Wainer 1993).

DIF analyses are designed to differentiate, across groups, between real differences in the construct being measured and sources of group-related construct-irrelevant variance. Different groups are not evaluated in terms of their differences in performance but rather in terms of differences in performance on each item, given the candidates' standings on the construct being measured, as indicated by the test taker's total score on the test (or some other relevant matching variable). DIF analyses provide an especially appealing way to address fairness issues, because the data required for DIF analyses (i.e., item responses and test scores) are readily available for most standardized testing programs and because DIF analyses provide a direct way to decrease construct-irrelevant differential impact (by avoiding the use of items with high DIF).

Zieky (2011) has provided a particularly interesting and informative analysis of the origins of DIF methodology. As noted earlier, from ETS's inception, its research staff had been concerned about fairness issues and had been actively investigating group differences in performance since the 1960s (Angoff and Ford 1973; Angoff and Sharon 1974; Cardall and Coffman 1964; Cleary 1968), but no fully adequate methodology for addressing group differences at the item level had been identified. The need to address the many obstacles facing the effective implementation of DIF was imposed on ETS researchers in the early 1980s:

> In 1984, ETS settled a lawsuit with the Golden Rule Insurance Company by agreeing to use raw differences in the percentages correct on an item in deciding on which items to include in a test to license insurance agents in Illinois; if two items were available that both met test specifications, the item with the smallest black-white difference in percentage correct was to be used; any difference in the percentages was treated as bias "even if it were caused by real and relevant differences between the groups in average knowledge of the tested subject." (Zieky 2011, p. 116)

The Golden Rule procedure was seen as causing limited harm in a minimum-competency licensing context but was seen as much more problematic in other contexts in which candidates would be ranked in terms of cognitive abilities or achievement, and concern grew that test quality would suffer if test developers were required to use only items "with the smallest raw differences in percent correct between Black and White test takers, regardless of the causes of these differences" (Zieky 2011, pp. 117–118):

> The goal was an empirical means of distinguishing between real group differences in the knowledge and skill measured by the test and unfair differences inadvertently caused by biased aspects of items. Test developers wanted help in ensuring that items were fair, but each method tried so far either had methodological difficulties or was too unwieldy to use on an operational basis with a wide variety of tests and several groups of test takers. The threat of legislation that would mandate use of the Golden Rule procedure for all tests further motivated ETS staff members to adopt a practical measure of DIF. (p. 118)

In response, ETS researchers (e.g., Dorans and Holland 1993; Holland and Thayer 1988) developed procedures that evaluated differential group performance, conditional on test takers' relative standing on the attribute of interest. The DIF methodology developed at ETS is now widely used in testing programs that aid in making high-stakes decisions throughout the world.

E. Burton and Burton (1993) found that the differences in scores across groups did not narrow substantially after the implementation of DIF analyses. Test items are routinely screened for sensitivity and other possible sources of differential functioning before administration, and relatively few items are flagged by the DIF statistics. As Zwick (2006) noted,

> even in the absence of evidence that it affects overall scores, … DIF screening is important as a precaution against the inclusion of unreasonable test content and as a source of information that can contribute to the construction of better tests in the future. (p. 668)

DIF screening addresses an issue that has to be confronted for psychometric and ethical reasons. That these checks on the quality of test items turn up relatively few cases of questionable item content is an indication that the item development and screening procedures are working as intended.

16.4.4 Identifying and Addressing Specific Threats to Fairness/Validity

As illustrated in the two previous subsections, much of the research on fairness at ETS, and more generally in the measurement research community, has focused on the identification and estimation of differential impact and potential bias in prediction and selection, a global issue, and on DIF, which addresses particular group-specific item effects that can generate adverse impact or bias. However, some researchers have sought to address other potential threats to fairness and, therefore, to validity.

Xi (2010) pointed out that fairness is essential to validity and validity is essential to fairness. If we define validity in terms of the appropriateness of proposed interpretations and uses of scores, and fairness in terms of the appropriateness of proposed interpretations and uses of scores across groups, then fairness would be a necessary condition for validity; if we define fairness broadly in terms of social justice, then validity would be a necessary condition for fairness. Either way, the two concepts are closely related; as noted earlier, Turnbull referred to validity as the "amoral brother" of fairness (Dorans, Chap. 7, this volume; Turnbull 1951).

Xi (2010) combined fairness and validity in a common framework by evaluating fairness as comparable validity across groups within the population of interest. She proposed to identify and evaluate any fairness-based objections to proposed interpretations and uses of the test scores as a *fairness argument* that would focus on whether an interpretation is equally plausible for different groups and whether the decision rules are appropriate for the groups. Once the inferences and assumptions

inherent in the proposed interpretation and use of the test scores have been specified, they can be evaluated in terms of whether they apply equally well to different groups. For example, it can be difficult to detect construct underrepresentation in a testing program by qualitatively evaluating how well the content of the test represents the content of a relevant domain, but empirical results indicating that there are substantial differences across groups in the relationship between performance on the test and more thorough measures of performance in the domain as a whole could raise serious questions about the representativeness of the test content. This argument-based approach can help to focus research on serious, specific threats to fairness/validity (Messick 1989).

Dorans and colleagues (Dorans, Chap. 7, this volume; Dorans and Holland 2000; Holland and Dorans 2006) have addressed threats to fairness/validity that can arise in scaling/equating test scores across different forms of a test:

> Scores on different forms or editions of a test that are supposed to be used interchangeably should be related to each other in the same way across different subpopulations. Score equity assessment (SEA) uses subpopulation invariance of linking functions across important subpopulations to assess the interchangeability of the scores. (Dorans, Chap. 7, this volume)

If the different forms of the test are measuring the same construct or combination of attributes in the different subpopulations, the equating function should not depend on the subpopulation on which it is estimated, and

> one way to demonstrate that two test forms are not equatable is to show that the equating functions used to link their scores are not invariant across different subpopulations of examinees. Lack of invariance in a linking function indicates that the differential difficulty of the two test forms is not consistent across different groups. (Dorans, Chap. 7, this volume)

SEA uses the invariance of the linking function across groups to evaluate consistency of the proposed interpretation of scores across groups and, thereby, to evaluate the validity of the proposed interpretation.

Mislevy et al. (2013) sought to develop systematic procedures for minimizing threats to fairness due to specific construct-irrelevant sources of variance in the assessment materials or procedures. To the extent that a threat to validity can be identified in advance or concurrently, the threat could be eliminated by suitably modifying the materials or procedures; for example, if it is found that English language learners have not had a chance to learn specific nontechnical vocabulary in a mathematics item, that vocabulary could be changed or the specific words could be defined. Mislevy et al. combined the general methodology of "universal design" with the ECD framework. In doing so, they made use of M. von Davier's (2008) general diagnostic model as a psychometric framework to identify specific requirements in test tasks. Willingham (1999) argued that test uses would be likely to be fairer across groups if "the implications of design alternatives are carefully examined at the outset" (p. 235) but recognized that this examination would be difficult to do "without much more knowledge of subgroup strengths and weaknesses… than is normally available" (p. 236). Mislevy et al. (2013) have been working to develop the kind of knowledge needed to build more fairness into testing procedures from the design stage.

16.5 Messick's Unified Model of Construct Validity

Samuel Messick spent essentially all of his professional life at ETS, and during his long and productive career, he made important contributions to many parts of test theory and to ETS testing programs, some of which were mentioned earlier. In this section, we focus on his central role in the development of the construct validity model and its transformation into a comprehensive, unified model of validity (Messick 1975, 1988, 1989). Messick's unified model pulled the divergent strands in validity theory into a coherent framework, based on a broad view of the meaning of test scores and the values and consequences associated with the scores, and in doing so, he gave the consequences of score use a prominent role.

Messick got his bachelor's degree in psychology and natural sciences from the University of Pennsylvania in 1951, and he earned his doctorate from Princeton University in 1954, while serving as an ETS Psychometric Fellow. His doctoral dissertation, "The Perception of Attitude Relationships: A Multidimensional Scaling Approach to the Structuring of Social Attitudes," reflected his dual interest in quantitative methods and in personality theory and social psychology. He completed postdoctoral fellowships at the University of Illinois, studying personality dynamics, and at the Menninger Foundation, where he did research on cognition and personality and received clinical training. He started as a full-time research psychologist at ETS in 1956, and he remained there until his death in 1998. Messick also served as a visiting lecturer at Princeton University on personality theory, abnormal psychology, and human factors between 1956 and 1958 and again in 1960–1961.

Messick completed his doctoral and postdoctoral work and started his career at ETS just as the initial version of construct validity was being developed (Cronbach and Meehl 1955). As noted, he came to ETS with a strong background in personality theory (e.g., see Messick 1956, 1972), where constructs play a major role, and a strong background in quantitative methods (e.g., see Gulliksen and Messick 1960; Messick and Abelson 1957; Schiffman and Messick 1963). Construct validity was originally proposed as a way to justify interpretations of test scores in terms of psychological constructs (Cronbach and Meehl 1955), and as such, it focused on psychological theory. Subsequently, Loevinger (1957) suggested that the construct model could provide a framework for all of validity, and Messick made this suggestion a reality. Between the late 1960s and the 1990s, he developed a broadly defined construct-based framework for the validation of test score interpretations and uses; his unified framework had its most complete statement in his validity chapter in the third edition of *Educational Measurement* (Messick 1989).

As Messick pursued his career, he maintained his dual interest in psychological theory and quantitative methods, applying this broad background to problems in educational and psychological measurement (Jackson and Messick 1965; Jackson et al. 1957; Messick and Frederiksen 1958; Messick and Jackson 1958; Messick and Ross 1962). He had close, long-term collaborations with a number of research psychologists (e.g., Douglas Jackson, Nathan Kogan, and Lawrence Stricker). His long-term collaboration with Douglas Jackson, whom he met while they were both postdoctoral fellows at the Menninger Foundation, and with whom he coauthored more than 25 papers and chapters (Jackson 2002), was particularly productive.

Messick's evolving understanding of constructs, their measurement, and their vicissitudes was, no doubt, strongly influenced by his background in social psychology and personality theory and by his ongoing collaborations with colleagues with strong substantive interest in traits and their roles in psychological theory. His work reflected an ongoing concern about how to differentiate between constructs (Jackson and Messick 1958; Stricker et al. 1969), between content and style (Jackson and Messick 1958; Messick 1962, 1991), and between constructs and potential sources of irrelevant variance (Messick 1962, 1964, 1981b; Messick and Jackson 1958).

Given his background and interests, it is not surprising that Messick became an "early adopter" of the construct validity model. Throughout his career, Messick tended to focus on two related questions: Is the test a good measure of the trait or construct of interest, and how can the test scores be appropriately used (Messick 1964, 1965, 1970, 1975, 1977, 1980, 1989, 1994a, b)? For measures of personality, he addressed the first of these questions in terms of "two critical properties for the evaluation of the purported personality measure … the measure's *reliability* and its *construct validity"* (Messick 1964, p. 111). Even in cases where the primary interest is in predicting behavior as a basis for decision making, and therefore, where it is necessary to develop evidence for adequate predictive accuracy, he emphasized the importance of evaluating the construct validity of the scores:

> Instead of talking about the reliability and construct validity (or even the empirical validity) of the *test* per se, it might be better to talk about the reliability and construct validity of the *responses* to the test, as summarized in a particular score, thereby emphasizing that these test properties are relative to the processes used by the subjects in responding. (Messick 1964, p. 112)

Messick also exhibited an abiding concern about ethical issues in research and practice throughout his career (Messick 1964, 1970, 1975, 1977, 1980, 1981b, 1988, 1989, 1998, 2000). In 1965, he examined some criticisms of psychological testing and discussed the possibilities for regulation and self-regulation for testing. He espoused "an 'ethics of responsibility,' in which pragmatic evaluations of the consequences of alternative actions form the basis for particular ethical decisions" (p. 140). Messick (1965) went on to suggest that policies based on values reflect and determine how we see the world, in addition to their intended regulatory effects, and he focused on "the value-laden nature of validity and fairness as psychometric concepts" (Messick 2000, p. 4) throughout his career. It is to this concern with meaning and values in measurement that we now turn.

16.5.1 Meaning and Values in Measurement

Messick was consistent in emphasizing ethical issues in testing, the importance of construct validity in evaluating meaning and ethical questions, and the need to consider consequences in evaluating test use: "But the ethical question of '*Should* these actions be taken?' cannot be answered by a simple appeal to empirical validity alone. The various social consequences of these actions must be contended with"

(Messick and Anderson 1970, p. 86). In 1975, Messick published a seminal paper that focused on meaning and values in educational measurement and explored the central role of construct-based analyses in analyzing meaning and in anticipating consequences. In doing so, he sketched many of the themes that he would subsequently develop in more detail. The paper (Messick 1975) was the published version of his presidential speech to Division 5 (Evaluation and Measurement) of the American Psychological Association. The title, "The Standard Problem: Meaning and Values in Measurement and Evaluation," indicates the intended breadth of the discussion and its main themes. As would be appropriate for such a speech, Messick focused on big issues in the field, and we will summarize five of these: (a) the central role of construct-based reasoning and analysis in validation, (b) the importance of ruling out alternate explanations, (c) the need to be precise about the intended interpretations, (d) the importance of consequences, and (e) the role of content-related evidence in validation.

First, Messick emphasized the central role of construct-based reasoning and analysis in validation. He started the paper by saying that any discussion of the meaning of a measure should center on construct validity as the "evidential basis" for inferring score meaning, and he associated construct validity with basic scientific practice:

> Construct validation is the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning. The problem of developing evidence to support an inferential leap from an observed consistency to a construct that accounts for that consistency is a generic concern of all science. (Messick 1975, p. 955)

A central theme in the 1975 paper is the interplay between theory and data. Messick suggested that, in contrast to concurrent, predictive, and content-based approaches to validation, each of which focused on a specific question, construct validation involves hypothesis testing and "all of the philosophical and empirical means by which scientific theories are evaluated" (p. 956). He wrote, "The process of construct validation, then, links a particular measure to a more general theoretical construct, usually an attribute or process or trait, that itself may be embedded in a more comprehensive theoretical network" (Messick 1975, p. 955). Messick took construct validation to define validation in the social sciences but saw education as slow in adopting this view. A good part of Messick's (1975) exposition is devoted to suggestions for why education had not adopted the construct model more fully by the early 1970s and for why that field should expand its view of validation beyond simple content and predictive interpretations. He quoted Loevinger (1957) to the effect that, from a scientific point of view, construct validity is validity, but he went further, claiming that content and criterion analyses are not enough, even for applied decision making, and that "the meaning of the measure must also be pondered in order to evaluate responsibly the possible consequences of the proposed use" (Messick 1975, p. 956). Messick was not so much suggesting the adoption of a particular methodology but rather encouraging us to think deeply about meanings and consequences.

Second, Messick (1975) emphasized the importance of ruling out alternate explanations in evaluation and in validation. He suggested that it would be effective and efficient to

> direct attention from the outset to vulnerabilities in the theory by formulating *counterhypotheses*, or plausible alternative interpretations of the observed consistencies. If repeated challenges from a variety of plausible rival hypotheses can be systematically discounted, then the original interpretation becomes more firmly grounded. (p. 956)

Messick emphasized the role of convergent/divergent analyses in ruling out alternative explanations of test scores.

This emphasis on critically evaluating proposed interpretations by empirically checking their implications was at the heart of Cronbach and Meehl's (1955) formulation of construct validity, and it reflects Popper's (1965) view that conjecture and refutation define the basic methodology of science. Messick's insistence on the importance of this approach probably originates less in the kind of philosophy of science relied on by Cronbach and Meehl and more on his training as a psychologist and on his ongoing collaborations with psychologists, such as Jackson, Kogan, and Stricker. Messick had a strong background in measurement and scaling theory (Messick and Abelson 1957), and he maintained his interest in these areas and in the philosophy of science throughout his career (e.g., see Messick 1989, pp. 21–34). His writings, however, strongly suggested a tendency to start with a substantive problem in psychology and then to bring methodology and "philosophical conceits" (Messick 1989) to bear on the problem, rather than to start with a method and look for problems to which it can be applied. For example, Messick (1984, 1989; Messick and Kogan 1963) viewed cognitive styles as attributes of interest and not simply as sources of irrelevant variance.

Third, Messick (1975) recognized the need to be precise about the intended interpretations of the test scores. If the extent to which the test scores reflect the intended construct, rather than sources of irrelevant variance, is to be investigated, it is necessary to be clear about what is and is not being claimed in the construct interpretation, and a clear understanding of what is being claimed helps to identify plausible competing hypotheses. For example, in discussing the limitations of a simple content-based argument for the validity of a dictated spelling test, Messick pointed out that

> the inference of inability or incompetence from the absence of correct performance requires the elimination of a number of plausible rival hypotheses dealing with motivation, attention, deafness, and so forth. Thus, a report of failure to perform would be valid, but one of inability to perform would not necessarily be valid. The very use of the term *inability* invokes constructs of attribute and process, whereas a content-valid interpretation would stick to the outcomes. (p. 960)

To validate, or evaluate, the interpretation and use of the test scores, it is necessary to be clear about the meanings and values inherent in that interpretation and use.

Fourth, Messick (1975) gave substantial attention to values and consequences and suggested that, in considering any test use, two questions needed to be considered:

> First, is the test any good as a measure of the characteristic it is interpreted to assess? Second, should the test be used for the proposed purpose? The first question is a technical and scientific one and may be answered by appraising evidence bearing on the test's psychometric properties, especially construct validity. The second question is an ethical one, and its answer requires an evaluation of the potential consequences of the testing in terms of social values. We should be careful not to delude ourselves that answers to the first question are also sufficient answers to the second (except of course when a test's poor psychometric properties preclude its use). (p. 960)

Messick saw meaning and values as intertwined: "Just as values play an important role in measurement, where meaning is the central issue, so should meaning play an important role in evaluation, where values are the central issue" (p. 962). On one hand, the meanings assigned to scores reflect the intended uses of the scores in making claims about test takers and in making decisions. Therefore the meanings depend on the values inherent in these interpretations and uses. On the other hand, an analysis of the meaning of scores is fundamental to an evaluation of consequences because (a) the value of an outcome depends in part on how it is achieved (Messick 1970, 1975) and (b) an understanding of the meaning of scores and the processes associated with performance is needed to anticipate unintended consequences as well as intended effects of score uses.

Fifth, Messick (1975) recognized that content representativeness is an important issue in test development and score interpretation, but that, in itself, it cannot establish validity. For one, content coverage is a property of the test:

> The major problem here is that content validity in this restricted sense is focused upon test *forms* rather than test *scores*, upon *instruments* rather than *measurements*. Inferences in educational and psychological measurement are made from scores, … and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. (p. 960)

Messick suggested that Loevinger's (1957) substantive component of validity, defined as the extent to which the construct to be measured by the test can account for the properties of the items included in the test, "involves a confrontation between content representativeness and response consistency" (p. 961). The empirical analyses can result in the exclusion of some items because of perceived defects, or these analyses may suggest that the conception of the trait and the corresponding domain may need to be modified:

> These analyses offer evidence for the substantive component of construct validity to the extent that the resultant content of the test can be accounted for by the theory of the trait (along with collateral theories of test-taking behavior and method distortion). (p. 961)

Thus the substantive component goes beyond traditional notions of content validity to incorporate inferences and evidence on response consistency as well as on the extent to which the response patterns are consistent with our understanding of the corresponding construct.

16.5.2 A Unified but Faceted Framework for Validity

Over the following decade, Messick developed his unified, construct-based conception of validity in several directions. In the third edition of *Educational Measurement* (Messick 1989), he proposed a very broad and open framework for validity as scientific inquiry. The framework allows for different interpretations at different levels of abstraction and generality, and it encourages the use of multiple modes of inquiry. It also incorporates values and consequences. Given the many uses of testing in our society and the many interpretations entailed by these uses, Messick's unified model inevitably became complicated, but he wanted to get beyond the narrow views of validation in terms of content-, criterion-, and construct-related evidence:

> What is needed is a way of cutting and combining validity evidence that forestalls undue reliance on selected forms of evidence, that highlights the important though subsidiary role of specific content- and criterion-related evidence in support of construct validity in testing applications, and that formally brings consideration of value implications and social consequences into the validity framework. (Messick 1989, p. 20)

Messick organized his discussion of the roles of different kinds of evidence in validation in a 2 × 2 table (see Fig. 16.1) that he had introduced a decade earlier (Messick 1980). The table has four cells, defined in terms of the function of testing (interpretation or use) and the justification for testing (evidence or consequences):

> The *evidential basis of test interpretation* is construct validity. The *evidential basis of test use* is also construct validity, but as buttressed by evidence for the relevance of the test to the specific applied purpose and for the utility of the testing in the applied setting. The *consequential basis of test interpretation* is the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the ideology in which the theory is embedded…. Finally, the *consequential basis of test use* is the appraisal of both potential and actual social consequences of the applied testing. (Messick 1989, p. 20, emphasis added)

Messick acknowledged that these distinctions were "interlocking and overlapping" (p. 20) and therefore potentially "fuzzy" (p. 20), but he found the distinctions and resulting fourfold classification to be helpful in structuring his description of the unified model of construct validity.

	Test Interpretation	**Test Use**
Evidential Basis	**Construct Validity**	**Construct Validity + Relevance/Utility**
Consequential Basis	**Value Implications**	**Social Consequences**

Fig. 16.1 Messick's facets of validity. From *Test Validity and the Ethics of Assessment* (p. 30, Research Report No. RR-79-10), by S. Messick, 1979, Princeton, NJ: Educational Testing Service. Copyright 1979 by Educational Testing Service. Reprinted with permission

16.5.3 The Evidential Basis of Test Score Interpretations

Messick (1989) began his discussion of the evidential basis of score interpretation by focusing on construct validity: "Construct validity, in essence, comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and relationships with other variables" (p. 34). Messick saw convergent and discriminant evidence as "overarching concerns" in discounting construct-irrelevant variance and construct underrepresentation. *Construct-irrelevant variance* occurs to the extent that test score variance includes "excess reliable variance that is irrelevant to the interpreted construct" (p. 34). *Construct underrepresentation* occurs to the extent that "the test is too narrow and fails to include important dimensions or facets of the construct" (p. 34).

Messick (1989) sought to establish the "trustworthiness" of the proposed interpretation by ruling out the major threats to this interpretation. The basic idea is to develop a construct interpretation and then check on plausible threats to this interpretation. To the extent that the interpretation survives all serious challenges (i.e., the potential sources of construct-irrelevant variance and construct underrepresentation), it can be considered trustworthy. Messick was proposing that strong interpretations (i.e., in terms of constructs) be adopted, but he also displayed a recognition of the essential limits of various methods of inquiry. This recognition is the essence of the constructive-realist view he espoused; our constructed interpretations are ambitious, but they are constructed by us, and therefore they are fallible. As he concluded,

> validation in essence is scientific inquiry into score meaning—nothing more, but also nothing less. All of the existing techniques of scientific inquiry, as well as those newly emerging, are fair game for developing convergent and discriminant arguments to buttress the construct interpretation of test scores. (p. 56)

That is, rather than specify particular rules or guidelines for conducting construct validations, he suggested broad scientific inquiry that could provide support for and illuminate the limitations of proposed interpretations and uses of test scores.

Messick suggested that construct-irrelevant variance and construct underrepresentation should be considered serious when they interfere with intended interpretations and uses of scores to a substantial degree. The notion of "substantial" in this context is judgmental and depends on values, but the judgments are to be guided by the intended uses of the scores. This is one way in which interpretations and meanings are not value neutral.

16.5.4 The Evidential Basis of Test Score Use

According to Messick (1989), construct validity provides support for test uses. However, the justification of test use also requires evidence that the test is appropriate for a particular applied purpose in a specific applied setting: "The construct validity of score interpretation undergirds *all* score-based inferences, not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores" (pp. 63–64). Messick rejected simple notions of content validity in terms of domain representativeness in favor of an analysis of the constructs associated with the performance domain. "By making construct theories of the performance domain and of its key attributes more explicit, however, test construction and validation become more rational, and the supportive evidence sought becomes more attuned to the inferences made" (p. 64). Similarly, Messick rejected the simple model of predictive validity in terms of a purely statistical relationship between test scores and criterion scores in favor of a construct-based approach that focuses on hypotheses about relationships between predictor constructs and criterion constructs: "There is simply no good way to judge the appropriateness, relevance, and usefulness of predictive inferences in the absence of evidence as to what the predictor and criterion scores mean" (p. 64). In predictive contexts, it is the relationship between the characteristics of test takers and their future performances that is of interest. The observed relationship between predictor scores and criterion scores provides evidence relevant to this hypothetical relationship, but it does not exhaust the meaning of that relationship.

In elaborating on the evidential basis of test use, Messick (1989) discussed a number of particular kinds of score uses (e.g., employment, selection, licensure), and a number of issues that would need to be addressed (e.g., curriculum, instructional, or job relevance or representativeness; test–criterion relationships; the utility of criteria; and utility and fairness in decision making), rather than relying on what he called ad hoc targets. He kept the focus on construct validation and suggested that "one should strive to maximize the meaningfulness of score interpretation and to minimize construct-irrelevant test variance. The resulting construct-valid scores then provide empirical components for rationally defensible prediction systems and rational components for empirically informed decision making" (p. 65). Messick (1989) was quite consistent in insisting on the primacy of construct interpretations in validity, even in those areas where empirical methods had tended to predominate. He saw the construct theory of domain performance as the basis for developing both the criterion and the predictor. Constructs provided the structure for validation and the glue that held it all together.

16.5.5 The Consequential Basis of Test Score Interpretation

Messick (1989) saw the consequential basis of test score interpretation as involving an analysis of the value implications associated with the construct label, with the construct theory, and with the general conceptual frameworks, or ideologies, surrounding the theory. In doing so, he echoed his earlier emphasis (Messick 1980) on the role of values in validity:

> Constructs are broader conceptual categories than are test behaviors and they carry with them into score interpretation a variety of value connotations stemming from at least three major sources: the evaluative overtones of the construct labels themselves; the value connotations of the broader theories or nomological networks in which constructs are embedded; and the value implications of still broader ideologies about the nature of humankind, society, and science that color our manner of perceiving and proceeding. (Messick 1989, p. 59)

Neither constructs nor the tests developed to estimate constructs are dictated by the data, as such. We make decisions about the kinds of attributes that are of interest to us, and these choices are based on the values inherent in our views.

Messick (1989) saw values as pervading and shaping the interpretation and use of test scores and therefore saw the evaluation of value implications as an integral part of validation:

> In sum, the aim of this discussion of the consequential basis of test interpretation was to raise consciousness about the pervasive consequences of value-laden terms (which in any event cannot be avoided in either social action or social science) and about the need to take both substantive aspects and value aspects of score meaning into account in test validation. (p. 63)

Under a constructive-realist model, researchers have to decide how to carve up and interpret observable phenomena, and they should be clear about the values that shape these choices.

16.5.6 The Consequential Basis of Test Score Use

The last cell (bottom right) of Messick's progressive matrix addresses the social consequences of testing as an "integral part of validity" (Messick 1989, p. 84). The validity of a testing program is to be evaluated in terms of how well the program achieves its intended function or purpose without undue negative consequences:

> Judging validity in terms of whether a test does the job it is employed to do … that is, whether it serves its intended function or purpose—requires evaluation of the intended or unintended social consequences of test interpretation and use. The appropriateness of the intended testing purpose and the possible occurrence of unintended outcomes and side effects are the major issues. (pp. 84–85)

The central question is whether the testing program achieves its goals well enough and at a low enough cost (in terms of negative consequences, anticipated and unanticipated) that it should be used.

Messick's (1989) discussion of the consequences of testing comes right after an extended discussion of criterion-related evidence and analyses of utility, in terms of a specific criterion in selection, and it emphasizes that such utility analyses are important, but they are not enough. The evaluation, or validation, of a test score use requires an evaluation of all major consequences of the testing program and not simply evidence that a particular criterion is being estimated and optimized:

> Even if adverse testing consequences derive from valid test interpretation and use, the appraisal of the functional worth of the testing in pursuit of the intended ends should take into account all of the ends, both intended and unintended, that are advanced by the testing application, including not only individual and institutional effects but societal or systemic effects as well. Thus, although appraisal of intended ends of testing is a matter of social policy, it is not only a matter of policy formation but also of policy evaluation that weighs all of the outcomes and side effects of policy implementation by means of test scores. Such evaluation of the consequences and side effects of testing is a key aspect of the validation of test use. (p. 85)

Messick used the term *functional worth* to refer to the extent that a testing program achieves its intended goals and is relatively free of unintended negative consequences. He seems to contrast this concept with *test validity*, which focuses on the plausibility of the proposed interpretation of the test scores. The approach is unified, but the analysis in terms of the progressive matrix is structured, complex, and nuanced.

Messick (1989) made several points about the relationship between validity and functional worth. First, to the extent that consequences are relevant to the evaluation of a testing program (in terms of either validity or functional worth), both intended and unintended consequences are to be considered. Second, consequences are relevant to the evaluation of test validity if they result from construct-irrelevant characteristics of the testing program. Third, if the unintended consequences cannot be traced to construct-irrelevant aspects of the testing program, the evaluation of consequences, intended and unintended, becomes relevant to the functional worth of the testing program, which is in Messick's progressive matrix "an aspect of the validation of test use" (p. 85). Messick's main concern in his discussion of functional worth was to emphasize that in evaluating such worth, it is necessary to evaluate unintended negative consequences as well as intended, criterion outcomes so as to further inform judgments about test use.

Construct meaning entered Messick's (1989) discussion of the consequential basis of test use in large part as a framework for identifying unintended consequences that merit further study:

> But once again, the construct interpretation of the test scores plays a facilitating role. Just as the construct meaning of the scores afforded a rational basis for hypothesizing predictive relationships to criteria, construct meaning provides a rational basis for hypothesizing potential testing outcomes and for anticipating possible side effects. That is, the construct theory, by articulating links between processes and outcomes, provides clues to possible effects. Thus, evidence of construct meaning is not only essential for evaluating the import of testing consequences, it also helps determine where to look for testing consequences. (pp. 85–86)

Messick's unified framework for validity encourages us to think broadly and deeply, in this case in evaluating unintended consequences. He encouraged the use of multiple value perspectives in identifying and evaluating consequences. The unified framework for validity incorporates evaluations of the extent to which test scores reflect the construct of interest (employing a range of empirical and conceptual methods) and an evaluation of the appropriateness of the construct measures for the use at hand (employing a range of values and criteria), but ultimately, questions about how and where tests are used are policy issues.

Messick (1989) summarized the evidential and consequential bases of score interpretation and use in terms of the four cells in his progressive matrix:

> The process of construct interpretation inevitably places test scores both in a theoretical context of implied relationships to other constructs and in a value context of implied relationships to good and bad valuations, for example, of the desirability or undesirability of attributes and behaviors. Empirical appraisals of the former substantive relationships contribute to an *evidential basis of test interpretation*, that is, to construct validity. Judgmental appraisals of the latter value implications provide a *consequential basis of test interpretation*.
>
> The process of test use inevitably places test scores both in a theoretical context of implied relevance and utility and in a value context of implied means and ends. Empirical appraisals of the former issues of relevance and utility, along with construct validity contribute to an *evidential basis for test use*. Judgmental appraisals of the ends a proposed test use might lead to, that is, of the potential consequences of a proposed use and of the actual consequences of applied testing, provide a *consequential basis for test use*. (p. 89)

The four aspects of the unified, construct-based approach to validation provide a comprehensive framework for validation, but it is a framework intended to encourage and guide conversation and investigation. It was not intended as an algorithm or a checklist for validation.

Messick's (1989) chapter is sometimes criticized for being long and hard to read, and it is in places, but this perception should not be so surprising, because he was laying out a broad framework for validation; making the case for his proposal; putting it in historical context; and, to some extent, responding to earlier, current, and imagined future critics—not a straightforward task. When asked about the intended audience for his proposed framework, he replied, "Lee Cronbach" (M. Zieky, personal communication, May 20, 2014). As is true in most areas of scientific endeavor, theory development is an ongoing dialogue between conjectures and data, between abstract principles and applications, and between scholars with evolving points of view.

16.5.7 Validity as a Matter of Consequences

In one of his last papers, Messick (1998) revisited the philosophical conceits of his 1989 chapter, and in doing so, he reiterated the importance of values and consequences for validity:

> What needs to be valid are the inferences made about score meaning, namely, the score interpretation and its action implications for test use. Because value implications both derive from and contribute to score meaning, different value perspectives may lead to different score implications and hence to different validities of interpretation and use for the same scores. (p. 37)

Messick saw construct underrepresentation and construct-irrelevant variance as serious threats to validity in all cases, but he saw them as especially serious if they led to adverse consequences:

> All educational and psychological tests underrepresent their intended construct to some degree and all contain sources of irrelevant variance. The details of this underrepresentation and irrelevancy are typically unknown to the test maker or are minimized in test interpretation and use because they are deemed to be inconsequential. If noteworthy adverse consequences occur that are traceable to these two major sources of invalidity, however, then both score meaning and intended uses need to be modified to accommodate these findings. (p. 42)

And he continued, "This is precisely why unanticipated consequences constitute an important form of validity evidence. Unanticipated consequences signal that we may have been incomplete or off-target in test development and, hence, in test interpretation and use" (p. 43). Levels of construct underrepresentation and construct-irrelevant variance that would otherwise be acceptable would become unacceptable if it were shown that they had serious negative consequences.

16.5.8 The Central Messages

Messick's (1975, 1980, 1981a, 1988, 1989, 1995) treatment of validity is quite thorough and complex, but he consistently emphasizes a few basic conclusions.

First, validity is a unified concept. It is "an integrated evaluative judgment" of the degree to which evidence and rationales support the inferences and actions based on test scores. We do not have "kinds" of validity for different score interpretations or uses.

Second, all validity is construct validity. Construct validity provides the framework for the unified model of validity because it subsumes both the content and criterion models and reflects the general practice of science in which observation is guided by theory.

Third, validation is scientific inquiry. It is not a checklist or procedure but rather a search for the meaning and justification of score interpretations and uses. The meaning of the scores is always important, even in applied settings, because meaning guides both score interpretation and score use. Similarly, values guide the construction of meaning and the goals of test score use.

Fourth, validity and science are value laden. Construct labels, theories, and supporting conceptual frameworks involve values, either explicitly or implicitly, and it is good to be clear about the underlying assumptions. It is better to be explicit than implicit about our values.

Fifth, Messick maintained that validity involves the appraisal of social consequences of score uses. Evaluating whether a test is doing what it was intended to do necessarily involves an evaluation of intended and unintended consequences.

There were two general concerns that animated Messick's work on validity theory over his career, both of which were evident from his earliest work to his last papers. One was his abiding interest in psychological theory and in being clear and explicit about the theoretical and pragmatic assumptions being made. Like Cronbach, he was convinced that we cannot do without theory and, more specifically, theoretical constructs, and rather than ignoring substantive, theoretical assumptions, he worked to understand the connections between theories, constructs, and testing.

The second was his abiding interest in values, ethics, and consequences, which was evident in his writing from the 1960s (Messick 1965) to the end of his career (Messick 1998). He recognized that values influence what we look at and what we see and that if we try to exclude values from our testing programs, we will tend to make the values implicit and unexamined. So he saw a role for values in evaluating the validity of both the interpretations of test scores and the uses of those scores. He did not advocate that the measurement community should try to impose any particular set of values, but he was emphatic and consistent in emphasizing that we should recognize and make public the value implications inherent in score interpretations and uses.

16.6 Argument-Based Approaches to Validation

Over a period of about 25 years, from the early 1960s to 1989, Messick developed a broad construct-based framework for validation that incorporated concerns about score interpretations and uses, meaning and values, scientific reasoning and ethics, and the interactions among these different components. As a result, the framework was quite complex and difficult to employ in applied settings.

Since the early 1990s, researchers have developed several related approaches to validation (Kane 1992, 2006, 2013a; Mislevy 2006, 2009; Mislevy et al. 1999; Mislevy et al. 2003b; Shepard 1993) that have sought to streamline models of validity and to add some more explicit guidelines for validation by stating the intended interpretation and use of the scores in the form of an argument. The argument would provide an explicit statement of the claims inherent in the proposed interpretation and use of the scores (Cronbach 1988).

By explicitly stating the intended uses of test scores and the score interpretations supporting these uses, these argument-based approaches seek to identify the kinds of evidence needed to evaluate the proposed interpretation and use of the test scores and thereby to specify necessary and sufficient conditions for validation.

Kane (1992, 2006) suggested that the proposed interpretation and use of test scores could be specified in terms of an interpretive argument. After coming to ETS, he extended the argument-based framework to focus on an interpretation/use argu-

ment (IUA), a network of inferences and supporting assumptions leading from a test taker's observed performances on test tasks or items to the interpretive claims and decisions based on the test scores (Kane 2013a). Some of the inferences in the IUA would be statistical (e.g., generalization from an observed score to a universe score or latent variable, or a prediction of future performance); other inferences would rely on expert judgment (e.g., scoring, extrapolations from the testing context to nontest contexts); and many of the inferences might be evaluated in terms of several kinds of evidence.

Most of the inferences in the IUA would be presumptive in the sense that the inference would establish a presumption in favor of its conclusion, or claim, but it would not prove the conclusion or claim. The inference could include qualitative qualifiers (involving words such as "usually") or quantitative qualifiers (e.g., standard errors or confidence intervals), as well as conditions under which the inference would not apply. The IUA is intended to represent the claims being made in interpreting and using scores and is not limited to any particular kind of claim.

The IUAs for most interpretations and uses would involve a chain of linked inferences leading from the test performances to claims based on these performances; the conclusion of one inference would provide the starting point, or datum, for subsequent inferences. The IUA is intended to provide a fairly detailed specification of the reasoning inherent in the proposed interpretation and uses of the test scores. Assuming that the IUA is coherent, in the sense that it hangs together, and complete, in the sense that it fully represents the proposed interpretation and use of the scores, it provides a clear framework for validation. The inferences and supporting assumptions in the IUA can be evaluated using evidence relevant to their plausibility. If all of the inferences and assumptions hold up under critical evaluation (conceptual and empirical), the interpretation and use can be accepted as plausible, or valid; if any of the inferences or assumptions fail to hold up under critical evaluation, the proposed interpretation and use of the scores would not be considered valid.

An argument-based approach provides a validation framework that gives less attention to philosophical foundations and general concerns about the relationship between meaning and values than did Messick's unified, construct-based validation framework, and more attention to the specific IUA under consideration. In doing so, an argument-based approach can provide necessary and sufficient conditions for validity in terms of the plausibility of the inferences and assumptions in the IUA. The validity argument is contingent on the specific interpretation and use outlined in the IUA; it is the proposed interpretation and uses that are validated and not the test or the test scores.

The argument-based approach recognizes the importance of philosophical foundations and of the relationship between meaning and values, but it focuses on how these issues play out in the context of particular testing programs with a particular interpretation and use proposed for the test scores. The conclusions of such argument-based analyses depend on the characteristics of the testing program and the proposed interpretation and uses of the scores; the claims being based on the test scores are specified and the validation effort is limited to evaluating these claims.

Chapelle et al. (2008, 2010) used the argument-based approach to analyze the validity of the *TOEFL*® test in some detail and, in doing so, provided insight into the meaning of the scores as well as their empirical characteristics and value implications. In this work, it is clear how the emphasis in the original conception of construct validity (Cronbach and Meehl 1955) on the need for a program of validation research rather than a single study and Messick's emphasis on the need to rule out threats to validity (e.g., construct-irrelevant variance and construct underrepresentation) play out in an argument-based approach to validation.

Mislevy (1993, 1994, 1996, 2007) focused on the role of evidence in validation, particularly in terms of model-based reasoning from observed performances to more general claims about students and other test takers. Mislevy et al. (1999, 2002, 2003a, b) developed an ECD framework that employs argument-based reasoning. ECD starts with an analysis of the attributes, or constructs, of interest and the social and cognitive contexts in which they function and then designs the assessment to generate the kinds and amounts of evidence needed to draw the intended inferences. The ECD framework involves several stages of analysis (Mislevy and Haertel 2006; Mislevy et al. 1999, 2002, 2003a). The first stage, *domain analysis*, concentrates on building substantive understanding of the performance domain of interest, including theoretical conceptions and empirical research on student learning and performance, and the kinds of situations in which the performances are likely to occur. The goal of this first stage is to develop an understanding of how individuals interact with tasks and contexts in the domain.

At the second stage, *domain modeling*, the relationships between student characteristics, task characteristics, and situational variables are specified (Mislevy et al. 2003a, b). The structure of the assessment to be developed begins to take shape, as the kinds of evidence that would be relevant to the goals of the assessment are identified.

The third stage involves the development of a *conceptual assessment framework* that specifies the operational components of the test and the relationships among these components, including a student model, task models, and evidence models. The student model provides an abstract account of the student in terms of ability parameters (e.g., in an IRT model). Task models posit schemas for collecting data that can be used to estimate the student parameters and guidelines for task development. The evidence model describes how student performances are to be evaluated, or scored, and how estimates of student parameters can be made or updated. With this machinery in place, student performances on a sample of relevant tasks can be used to draw probabilistic inferences about student characteristics.

The two dominant threads in these argument-based approaches to validation are the requirement that the claims to be made about test takers (i.e., the proposed interpretation and use of the scores) be specified in advance, and then justified, and that inferences about specific test takers be supported by warrants or models that have been validated, using empirical evidence and theoretical rationales. The argument-based approaches are consistent with Messick's unified framework, but they tend to focus more on specific methodologies for the validation of proposed interpretations and uses than did the unified framework.

16.7 Applied Validity Research at ETS

In addition to the contributions to validity theory described above, ETS research has addressed numerous practical issues in documenting the validity of various score uses and interpretations and in identifying the threats to the validity of ETS tests. Relatively straightforward predictive validity studies were conducted at ETS from its earliest days, but ETS research also has addressed problems in broadening both the predictor and criterion spaces and in finding better ways of expressing the results of predictive validity studies. Samuel Messick's seminal chapter in the third edition of *Educational Measurement* (Messick 1989) focused attention on the importance of identifying factors contributing to construct-irrelevant variance and identifying instances of construct underrepresentation, and numerous ETS studies have focused on both of these problems.

16.7.1 Predictive Validity

Consistent with the fundamental claim that tests such as the SAT test were useful because they could predict academic performance, predictive validity studies were common throughout the history of ETS. As noted earlier, the second Research Bulletin published by ETS (RB-48-02) was a predictive study titled *The Prediction of First Term Grades at Hamilton College* (Frederiksen 1948). The abstract noted, "It was found that the best single predictor of first term average grade was rank in secondary school (r = .57). The combination of SAT scores with school rank was found to improve the prediction considerably (R = .67)." By 1949, enough predictive validity studies had been completed that results of 17 such studies could be summarized by Allen (1949). This kind of study was frequently repeated over the years, but even in the very earliest days there was considerable attention to a more nuanced view of predictive validity from both the perspective of potential predictors and potential criteria. As noted, the Frederiksen study cited earlier was the *second* Research Bulletin published by ETS, but the *first* study published (College Board 1948) examined the relationship of entrance test scores at the U.S. Coast Guard Academy to outcome variables that included both course grades and nonacademic ratings. On the predictor side, the study proposed that "a cadet's standing at the Academy be based on composite scores based on three desirable traits: athletic ability, adaptability, and academic ability." A follow-up study (French 1948) included intercorrelations of 76 measures that included academic and nonacademic tests as predictors and included grades and personality ratings as criteria. The conclusions supported the use of the academic entrance tests but noted that the nonacademic tests in that particular battery did not correlate with either grades or personality ratings.

Although there were a number of studies focusing on the prediction of first-year grades in the 1950s (e.g., Abelson 1952; Frederiksen et al. 1950a, b; Mollenkopf

1951; Schultz 1952), a number of studies went beyond that limited criterion. For example, Johnson and Olsen (1952) compared 1-year and 3-year predictive validities of the Law School Admissions Test in predicting grades. Mollenkopf (1950) studied the ability of aptitude and achievement tests to predict both first- and second-year grades at the U.S. Naval Postgraduate School. Although the second-year validities were described as "fairly satisfactory," they were substantially lower than the Year 1 correlations. This difference was attributed to a number of factors, including differences in the first- and second-year curricula, lower reliability of second-year grades, and selective dropout. Besides looking beyond the first year, these early studies also considered other criteria. French (1957), in a study of 12th-grade students at 42 secondary schools, related SAT scores and scores on the Tests of Developed Ability (TDA) to criteria that included high school grades but also included students' self-reports of their experiences and interests and estimations of their own abilities. In addition, teachers nominated students who they believed exhibited outstanding ability. The study concluded not only that the TDA predicted grades in physics, chemistry, biology, and mathematics but that, more so than the SAT, it was associated with self-reported scientific interests and experiences.

From the 1960s through the 1980s, ETS conducted a number of SAT validity studies that focused on routine predictions of the freshman grade point average (FGPA) with data provided from colleges using the College Board/ETS Validity Study Service as summarized by Ramist and Weiss (1990). In 1994, Ramist et al. (1994) produced a groundbreaking SAT validity study that introduced a number of innovations not found in prior work. First, the study focused on course grades, rather than FGPA, as the criterion. Because some courses are graded much more strictly than others, when grades from these courses are combined without adjustment in the FGPA, the ability of the SAT to predict freshman performance is underestimated. Several different ways of making the adjustment were described and demonstrated. Second, the study corrected for the range restriction in the predictors caused by absence of data for the low-scoring students not admitted to college. (Although the range restriction formulas were not new, they had not typically been employed in multicollege SAT validity studies.) Third, the authors adjusted course grades for unreliability. Fourth, they provided analyses separately for a number of subgroups defined by gender, ethnicity, best language, college selectivity, and college size. When adjustments were made for multivariate range restriction in the predictors, grading harshness/leniency for specific courses, and criterion unreliability, the correlation of the SAT with the adjusted grades was .64, and the multiple correlation of SAT and high school record with college grades was .75.

Subsequent SAT validity studies incorporated a number of these methods and provided new alternatives. Bridgeman et al. (2000), for example, used the course difficulty adjustments from the 1994 study but noted that the adjustments could be quite labor intensive for colleges trying to conduct their own validity studies. They showed that simply dividing students into two categories based on intended major (math/science [where courses tend to be severely graded] vs. other) recovered many of the predictive benefits of the complex course difficulty adjustments. In a variation on this theme, a later study by Bridgeman et al. (2008c) provided correlations sepa-

rately for courses in four categories (English, social science, education, and science/math/engineering) and focused on cumulative grades over an entire college career, not just the first year. This study also showed that, contrary to the belief that the SAT predicts only FGPA, predictions of cumulative GPA over 4 or 5 years are similar to FGPA predictions.

16.7.2 Beyond Correlations

From the 1950s through the early 2000s, the predictive validity studies for the major admissions testing programs (e.g., SAT, the *GRE*® test, GMAT) tended to rely on correlations to characterize the relationship between test scores and grades. Test critics would often focus on unadjusted correlations (typically around .30). Squaring this number to get "variance accounted for," the critics would suggest that a test that explained less than 10% of the variance in grades must be of very little practical value (e.g., Fairtest 2003). To counter this perception, Bridgeman and colleagues started supplementing correlational results by showing the percentage of students who would succeed in college at various score levels (e.g., Bridgeman et al. 2008a, b, c; Cho and Bridgeman 2012). For example, in one study, 12,529 students at moderately selective colleges who had high school GPAs of at least 3.7 were divided into groups based on their combined Verbal and Mathematics SAT scores (Bridgeman et al. 2008a). Although college success can be defined in many different ways, this study defined success relatively rigorously as achieving a GPA of 3.5 or higher at the end of the college career. For students with total SAT scores (verbal + mathematics) of 1000 or lower, only 22% had achieved this level of success, whereas 73% of students in the 1410–1600 score category had finished college with a 3.5 or higher. Although SAT scores explained only about 12% of the variance in the overall group (which may seem small), the difference between 22% and 73% is substantial. This general approach to meaningful presentation of predictive validity results was certainly not new; rather, it is an approach that must be periodically rediscovered. As Ben Schrader noted in 1965,

> during the past 60 years, correlation and regression have come to occupy a central position in measurement and research…. Psychologists and educational researchers use these methods with confidence based on familiarity. Many persons concerned with research and testing, however, find results expressed in these terms difficult or impossible to interpret, and prefer to have results expressed in more concrete form. (p. 29)

He then went on to describe a method using expectancy tables that showed how standing on the predictor, in terms of fifths, related to standing on the criterion, also in terms of fifths. He used scores on the Law School Admission Test as the predictor and law school grades as the criterion. Even the 1965 interest in expectancy tables was itself a rediscovery of their explanatory value. In their study titled "Prediction of First Semester Grades at Kenyon College, 1948–1949," Frederiksen et al. (1950a) included an expectancy table that showed the chances in 100 that a student would

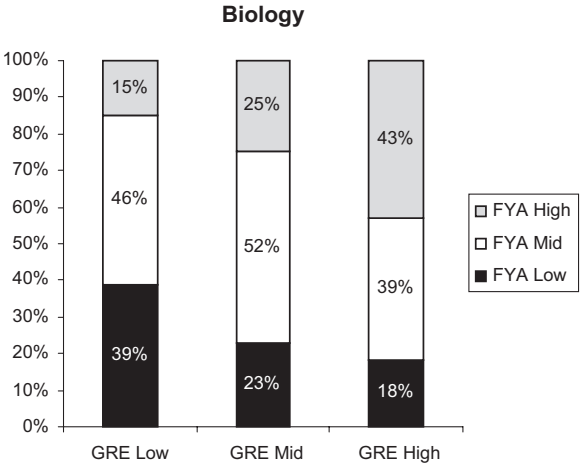

Fig. 16.2 Percentage of biology graduate students in GRE quartile categories whose graduate grade point averages were in the bottom quartile, mid-50%, or top quartile of the students in their departments across 24 programs (Adapted from Bridgeman et al. 2008a. Copyright 2008 by Educational Testing Service. Used with permission)

earn an average of at least a specified letter grade given a predicted grade based on a combination of high school rank and SAT scores. For example, for a predicted grade of B, the chance in 100 of getting at least a C+ was 88, at least a B was 50, and at least an A− was 12.

Despite the appeal of the expectancy table approach, it lay dormant until modest graphical extensions of Schrader's ideas were again introduced in 2008 and beyond. An example of this more graphical approach is in Fig. 16.2 (Bridgeman et al. 2008a, p. 10). Within each of 24 graduate biology programs, students were divided into quartiles based on graduate grades and into quartiles based on combined GRE verbal and quantitative scores. These results were then aggregated across the 24 programs and graphed. The graph shows that almost three times as many students with top quartile GRE scores were in the high-GPA category (top quartile) compared to the number of high-GPA students in the bottom GRE quartile.

The same report also used a graphical approach to show a kind of incremental validity information. Specifically, the bottom and top quartiles in each department were defined in terms of both undergraduate grade point average (UGPA) and GRE scores. Then, within the bottom UGPA quartile, students with top or bottom GRE scores could be compared (and similarly for the top UGPA quartile). Because graduate grades tend to be high, success was defined as achieving a 4.0 grade average. Figure 16.3 indicates that, even within a UGPA quartile, GRE scores matter for identifying highly successful students (i.e., the percentage achieving a 4.0 average).

16.7.3 Construct-Irrelevant Variance

The construct-irrelevant factors that can influence test scores are almost limitless. A comprehensive review of all ETS studies related to construct-irrelevant variance would well exceed the space limitations in this document; rather, a sampling of

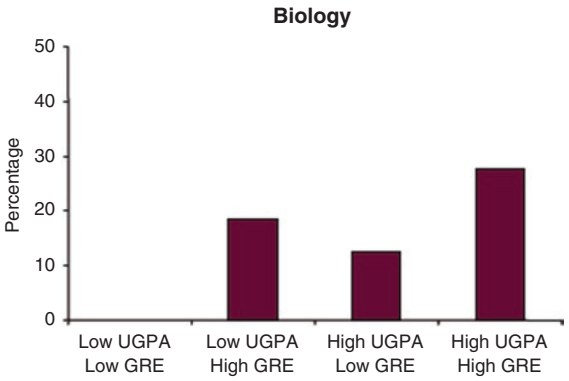

Fig. 16.3 Percentage of students in graduate biology departments earning a 4.0 grade point average by undergraduate grade point average and GRE high and low quartiles (Adapted from Bridgeman et al. 2008a Copyright 2008 by Educational Testing Service. Used with permission)

studies that explore various aspects of construct-irrelevant variance is presented. Research on one source of irrelevant variance, coaching, is described in a separate chapter by Donald Powers (Chap. 17, this volume).

16.7.3.1 Fatigue Effects

The potential for test-taker fatigue to interfere with test scores was already a concern in 1948, as suggested by the title of ETS Research Memorandum No. 48-02 by Tucker (1948), *Memorandum Concerning Study of Effects of Fatigue on Afternoon Achievement Test Scores Due to Scholastic Aptitude Test Being Taken in the Morning*. A literature review on the effects of fatigue on test scores completed in 1966 reached three conclusions:

> 1) Sufficient evidence exists in the literature to discount any likelihood of physiological consequences to the development of fatigue during a candidate's taking the College Board SAT or Achievement Tests; 2) the decline in feeling-tone experienced by an individual is often symptomatic of developing fatigue, but this decline does not necessarily indicate a decline in the quantity or quality of work output; and 3) the amount of fatigue that develops as a result of mental work is related to the individual's conception of, and attitude and motivation toward, the task being performed. (Wohlhueter 1966, Abstract)

A more recent experimental study conducted when the SAT was lengthened by the addition of the writing section reached a similar conclusion: "Results indicated that while the extended testing time for the new SAT may cause test takers to feel fatigued, fatigue did not affect test taker performance" (Liu et al. 2004, Abstract).

16.7.3.2 Time Limits

If a test is designed to assess speed of responding, then time limits merely enforce construct-relevant variance. But if the time limit is imposed primarily for administrative convenience, then a strict time limit might not be construct relevant. On one hand, an early study on the influence of timing on Cooperative Reading Test scores

suggested no significant changes in means or standard deviations with extended time (Frederiksen 1951). On the other hand, Lord (1953, Abstract) concluded that "unspeeded (power) tests are more valid" based on a study of 649 students at one institution. Evans (1980) created four SAT-like test forms that were administered in one of three speededness conditions: normal, speeded, and unspeeded. Degree of speededness affected scores but did not interact with gender or ethnicity. The technical handbook for the SAT by Donlon (1984) indicated that the speed with which students can answer the questions should play only a very minor role in determining scores. A study of the impact of extending the amount of time allowed per item on the SAT concluded that there were some effects of extended time (1.5 times regular time); average gains for the verbal score were less than 10 points on the 200–800 scale and about 30 points for the mathematics scores (Bridgeman et al. 2004b). But these effects varied considerably depending on the ability level of the test taker. Somewhat surprisingly, for students with SAT scores of 400 or lower, extra time had absolutely no impact on scores. Effects did not interact with either gender or ethnicity. Extended time on the GRE was similarly of only minimal benefit with an average increase of 7 points for both verbal and quantitative scores on the 200–800 scale when the time limit was extended to 1.5 times standard time (Bridgeman et al. 2004a).

When new tests are created or existing tests are modified, appropriate time limits must be set. A special timing study was conducted when new item types were to be introduced to the SAT to provide an estimate of the approximate amount of time required to answer new and existing item types (Bridgeman and Cahalan 2007). The study used three approaches to estimate the amount of time needed to answer questions of different types and difficulties: (a) Item times were automatically recorded from a computer-adaptive version of the SAT, (b) students were observed from behind a one-way mirror in a lab setting as they answered SAT questions under strict time limits and the amount of time taken for each question was recorded, and (c) high school students recorded the amount of time taken for test subsections that were composed of items of a single type. The study found that the rules of thumb used by test developers were generally accurate in rank ordering the item types from least to most time consuming but that the time needed for each question was higher than assumed by test developers.

Setting appropriate time limits that do not introduce construct-irrelevant variance is an especially daunting challenge for evaluating students with disabilities, as extended time is the most common accommodation for these students. Evaluating the appropriateness of extended time limits for students with disabilities has been the subject of several research reports (e.g., Cahalan et al. 2006; Packer 1987; Ragosta and Wendler 1992) as well as receiving considerable attention in the book *Testing Handicapped People* (Willingham et al. 1988).

Setting appropriate time limits on a computer-adaptive test (CAT) in which different students respond to different items can be especially problematic. Bridgeman and Cline (2000) showed that when the GRE was administered as a CAT, items at the same difficulty level and meeting the same general content specifications could vary greatly in the time needed to answer them. For example, a question assessing

the ability to add numbers with negative exponents could be answered very quickly while a question at the same difficulty level that required the solution of a pair of simultaneous equations would require much more time even for very able students. Test takers who by chance received questions that could be answered quickly would then have an advantage on a test with relatively strict time limits. Furthermore, running out of time on a CAT and guessing to avoid the penalty for an incomplete test can have a substantial impact on the test score because the CAT scoring algorithm assumed that an incorrect answer reflected a lack of ability and not an unlucky guess (Bridgeman and Cline 2004). A string of unlucky guesses at the end of the GRE CAT (because the test taker ran out of time and had to randomly respond) could lower the estimated score by more than 100 points (on a 200–800 scale) compared to the estimated score when the guessing began.

16.7.3.3 Guessing

Guessing can be a source of construct-irrelevant variance because noise is added to measurement precision when test takers answer correctly by guessing but actually know nothing about the answer (Wendler and Walker 2006). Corrections for guessing often referred to as formula scoring attempt to limit this irrelevant variance by applying a penalty for incorrect answers so that answering incorrectly has more negative consequences than merely leaving a question blank. For example, with the five-option multiple-choice questions on the SAT (prior to 2016), a test taker received 1 point for a correct answer and 0 points for an omitted answer, and one-fourth of a point was subtracted for each incorrect answer. (The revised SAT introduced in 2016 no longer has a correction for guessing.) By the time ETS was founded, there were already more than 20 years of research on the wisdom and effects of guessing corrections. Freeman (1952) surveyed this research and observed,

> At the outset, it may be stated that the evidence is not conclusive. While much that is significant has been written about the theoretical need to correct for guessing, and about the psychological and instructional value of such a correction, the somewhat atomistic, or at least uncoordinated, research that has been done during the last 25 years fails to provide an answer that can be generalized widely. (p. 1)

More than 60 years later, research is still somewhat contradictory and a definitive answer is still illusive. Lord (1974) argued that under certain assumptions, formula scoring is "clearly superior" to number-right scoring, though it remains unclear how often those assumptions are actually met. Angoff (1987) conducted an experimental study with different guessing instructions for SAT Verbal items and concluded, "Formula scoring is not disadvantageous to students who are less willing to guess and attempt items when they are not sure of the correct answer" (abstract). Conversely, some individuals and population subgroups may differ in their willingness to guess so that conclusions based on averages in the population as a whole may not be valid for all people. Rivera and Schmitt (1988), for example, noted a difference in willingness to guess on the part of Hispanic test takers, especially

Mexican Americans. Beginning in the 1981–1982 test year, the GRE General Test dropped formula scoring and became a rights-only scored test, but the GRE Subject Tests retained formula scoring. In 2011, the *Advanced Placement*® (*AP*®) test program dropped formula scoring and the penalty for incorrect answers. At the end of 2014, the SAT was still using formula scoring, but the announcement had already been made that the revised SAT would use rights-only scoring.

16.7.3.4 Scoring Errors

Any mistakes made in scoring a test will contribute to irrelevant variance. Although the accuracy of machine scoring of multiple-choice questions is now almost taken for granted, early in the history of ETS, there were some concerns with the quality of the scores produced by the scanner. Note that the formula scoring policy put special demands on the scoring machine because omitted answers and incorrect answers were treated differently. The machine needed to determine if a light mark was likely caused by an incomplete erasure (indicating intent to omit) or if the relatively light mark was indeed the intended answer. The importance of the problem may be gauged by the status of the authors, Fan, Lord, and Tucker, who devised "a system for reducing the number of errors in machine-scoring of multiple-choice answer sheets" (Fan et al. 1950, Abstract). Measuring and reducing rater-related scoring errors on essays and other constructed responses were also of very early concern. A study of the reading reliability of the College Board English Composition test was completed in 1948 (Aronson 1948; ETS 1948). In the following years, controlling irrelevant variance introduced by raters of constructed responses (whether human or machine) was the subject of a great deal of research, which is discussed in another chapter (Bejar, Chap. 18, this volume).

16.7.4 Construct Underrepresentation

Whereas construct-irrelevant variance describes factors that should not contribute to test scores, but do, construct underrepresentation is the opposite—failing to include factors in the assessment that should contribute to the measurement of a particular construct. If the purpose of a test or battery of tests is to assess the likelihood of success in college (i.e., the construct of interest), failure to measure the noncognitive skills that contribute to such success could be considered a case of construct underrepresentation. As noted, from the earliest days of ETS, there was interest in assessing more than just verbal and quantitative skills. In 1948, the organization's first president, Chauncey, called for a "Census of Abilities" that would assess attributes that went beyond just verbal and quantitative skills to include "personal qualities, … drive (energy), motivation (focus of energy), conscientiousness, … ability to get along with others" (Lemann 1995, p. 84). From 1959 to 1967, ETS had a personality research group headed by Samuel Messick. The story of personality

research at ETS is described in two other chapters (Kogan, Chap. 14, this volume; Stricker, Chap. 13, this volume).

Despite the apparent value of broadening the college-readiness construct beyond verbal and quantitative skills, the potential of such additional measures as a part of operational testing programs needed to be rediscovered from time to time. Frederiksen and Ward (1978) described a set of tests of scientific thinking that were developed as potential criterion measures, though they could also be thought of as additional predictors. The tests assessed both quality and quantity of ideas in formulating hypotheses and solving methodological problems. In a longitudinal study of 3,500 candidates for admission to graduate programs in psychology, scores were found to be related to self-appraisals of professional skills, professional accomplishments in collaborating in research, designing research apparatus, and publishing scientific papers. In a groundbreaking article in the *American Psychologist*, Norman Frederiksen (1984) expanded the argument for a broader conception of the kinds of skills that should be assessed. In the article, titled "The Real Test Bias: Influences of Testing on Teaching and Learning," Frederiksen argued that

> there is evidence that tests influence teacher and student performance and that multiple-choice tests tend not to measure the more complex cognitive abilities. The more economical multiple-choice tests have nearly driven out other testing procedures that might be used in school evaluation. (Abstract)

Another article, published in the same year, emphasized the critical role of social intelligence (Carlson et al. 1984). The importance of assessing personal qualities in addition to academic ability for predicting success in college was further advanced in a multiyear, multicampus study that was the subject of two books (Willingham 1985; Willingham and Breland 1982). This study indicated the importance of expanding both the predictor and criterion spaces. The study found that if the only criterion of interest is academic grades, SAT scores and high school grades appear to be the best available predictors, but, if criteria such as leadership in school activities or artistic accomplishment are of interest, the best predictors are previous successes in those areas.

Baird (1979) proposed a measure of documented accomplishments to provide additional evidence for graduate admissions decisions. In contrast to a simple listing of accomplishments, documented accomplishments require candidates to provide verifiable evidence for their claimed accomplishments. The biographical inventory developed in earlier stages was evaluated in 26 graduate departments that represented the fields of English, biology, and psychology. Responses to the inventory were generally not related to graduate grades, but a number of inventory responses reflecting preadmission accomplishments were significantly related to accomplishments in graduate school (Baird and Knapp 1981). Lawrence Stricker and colleagues further refined measures of documented accomplishments (Stricker et al. 2001).

Moving into the twenty-first century, there was rapidly increasing interest in noncognitive assessments (Kyllonen 2005), and a group was established at ETS to deal specifically with these new constructs (or to revisit older noncognitive constructs that in earlier years had failed to gain traction in operational testing programs). The label "noncognitive" is not really descriptive and was a catch-all that included any assessment that went beyond the verbal, quantitative, writing, and subject matter skills and knowledge that formed the backbone of most testing programs at ETS. Key noncognitive attributes include persistence, dependability, motivation, and teamwork. One measure that was incorporated into an operational program was the *ETS*® Personal Potential Index (*ETS*® PPI) service, which was a standardized rating system in which individuals who were familiar with candidates for graduate school, such as teachers or advisors, could rate core personal attributes: knowledge and creativity, resilience, communication skills, planning and organization, teamwork, and ethics and integrity. All students who registered to take the GRE were given free access to the PPI and a study was reported that demonstrated how the diversity of graduate classes could be improved by making the PPI part of the selection criteria (Klieger et al. 2013). Despite its potential value, the vast majority of graduate schools were reluctant to require the PPI, at least in part because they were afraid of putting in place any additional requirements that they thought might discourage applicants, especially if their competition did not have a similar requirement. Because of this very low usage, ETS determined that the resources needed to support this program could be better used elsewhere and, in 2015, announced the end of the PPI as part of the GRE program. This announcement certainly did not signal an end to interest in noncognitive assessments. A noncognitive assessment, the *SuccessNavigator*® assessment, which was designed to assist colleges in making course placement decisions, was in use at more than 150 colleges and universities in 2015. An ongoing research program provided evidence related to placement validity claims, reliability, and fairness of the measure's scores and placement recommendations (e.g., Markle et al. 2013; Rikoon et al. 2014).

The extent to which writing skills are an important part of the construct of readiness for college or graduate school also has been of interest for many years. Although a multiple-choice measure of English writing conventions, the Test of Standard Written English, was administered along with the SAT starting in 1977, it was seen more as an aid to placement into English classes than as part of the battery intended for admissions decisions. Rather than the 200–800 scale used for Verbal and Mathematics tests, it had a truncated scale running from 20 to 60. By 2005, the importance of writing skills to college preparedness was recognized by inclusion of a writing score based on both essay and multiple-choice questions and reported on the same 200–800 scale as Verbal and Mathematics. Starting in the mid-1990s, separately scored essay-based writing sections became a key feature of high-stakes admissions tests at ETS, starting with the GMAT, then moving on to the GRE and the *TOEFL iBT*® test. A major reason for the introduction of TOEFL iBT in 2005 was to broaden the academic English construct assessed (i.e., reduce the construct

underrepresentation) by adding sections on speaking and writing skills. By 2006, the *TOEIC*® tests, which are designed to evaluate English proficiency in the workplace, were also offering an essay section.

The importance of writing in providing adequate construct representation was made clear for AP tests by the discovery of nonequivalent gender differences on the multiple-choice and constructed-response sections of many AP tests (Mazzeo et al. 1993). That finding meant that a different gender mix of students would be granted AP credit depending on which item type was given more weight, including if only one question type was used. Bridgeman and Lewis (1994) noted that men scored substantially higher than women (by about half of a standard deviation) on multiple-choice portions of AP history examinations but that women and men scored almost the same on the essays and that women tended to get slightly higher grades in their college history courses. Furthermore, the composite of the multiple-choice and essay sections provided better prediction of college history grades than either section by itself for both genders. Thus, if the construct were underrepresented by a failure to include the essay section, not only would correlations have been lower but substantially fewer women would have been granted AP credit. Bridgeman and McHale (1998) performed a similar analysis for the GMAT, demonstrating that the addition of the essay would create more opportunities for women.

16.8 Fairness as a Core Concern in Validity

Fairness is a thread that has run consistently through this chapter because, as Turnbull (1951) and others have noted, the concepts of fairness and validity are very closely related. Also noted at a number of points in this chapter, ETS has been deeply concerned about issues of fairness and consequences for test takers as individuals throughout its existence, and these concerns have permeated its operational policies and its research program (Bennett, Chap. 1, this volume; Messick 1975, 1989, 1994a, 1998, 2000; Turnbull 1949, 1951). However, with few exceptions, measurement professionals paid little attention to fairness across groups until the 1960s (D.R. Green 1982), when this topic became a widespread concern among test developers and many test publishers instituted fairness reviews and empirical analyses to promote item and test fairness (Zieky 2006).

Messick's (1989) fourfold analysis of the evidential and consequential bases of test score interpretations and uses gave a lot of attention to evaluations of the fairness and overall effectiveness of testing programs in achieving intended outcomes and in minimizing unintended negative consequences. As indicated earlier, ETS researchers have played a major role in developing statistical models and methodology for identifying and controlling likely sources of construct-irrelevant variance and construct underrepresentation and thereby promoting fairness and reducing bias. In doing so, they have tried to clarify how the evaluation of consequences fits into a more general validation framework.

Frederiksen (1984, 1986) made the case that objective (multiple-choice) formats tended to measure a subset of the skills important for success in various contexts but that reliance on that format could have a negative, distorting effect on instruction. He recalled that, while conducting validity studies during the Second World War, he was surprised that reading comprehension tests and other verbal tests were the best predictors of grades in gunner's mate school. When he later visited the school, he found that the instruction was mostly lecture–demonstration based on the content of manuals, and the end-of-course tests were based on the lectures and manuals. Frederiksen's group introduced performance tests that required students to service real guns, and grades on the end-of-course tests declined sharply. As a result, the students began assembling and disassembling guns, and the instructors "moved out the classroom chairs and lecture podium and brought in more guns and gunmounts" (Frederiksen 1984, p. 201). Scores on the new performance tests improved. In addition, mechanical aptitude and knowledge became the best predictors of grades:

> No attempt was made to change the curriculum or teacher behavior. The dramatic changes in achievement came about solely through a change in the tests. The moral is clear: It is possible to influence teaching and learning by changing the tests of achievement. (p. 201)

Testing programs can have dramatic systemic consequences, positive or negative.

Negative consequences count against a decision rule (e.g., the use of a cut score), but they can be offset by positive consequences. A program can have substantial negative consequences and still be acceptable, if the benefits outweigh those costs. Negative consequences that are not offset by positive consequences tend to render a decision rule unacceptable (at least for stakeholders who are concerned about these consequences).

In reviewing a National Academy of Sciences report on ability testing (Wigdor and Garner 1982), Messick (1982b) suggested that the report was dispassionate and wise but that it "evinces a pervasive institutional bias" (p. 9) by focusing on common analytic models for selection and classification, which emphasize the intended outcomes of the decision rule:

> Consider that, for the most part, the utility of a test for selection is appraised statistically in terms of the correlation coefficient between the test and the criterion … but this correlation is directly proportional to the obtained gains over random selection in the criterion performance of the selected group…. Our traditional statistics tend to focus on the accepted group and on minimizing the number of poor performers who are accepted, with little or no attention to the rejected group or those rejected individuals who would have performed adequately if given the chance. (p. 10)

Messick went on to suggest that "by giving primacy to productivity and efficiency, the Committee simultaneously downplays the significance of other important goals in education and the workplace" (p. 11). It is certainly appropriate to evaluate a decision rule in terms of the extent to which it achieves the goals of the program, but it is also important to attend to unintended effects that have potentially serious consequences.

Holland (1994) and Dorans (2012) have pointed out that that different stakeholders (test developers, test users, test takers) can have very different but legitimate

perspectives on testing programs and on the criteria to be used in evaluating the programs. For some purposes and in some contexts, it is appropriate to think of testing programs primarily as measurement procedures designed to produce accurate and precise estimates of some variable of interest; within this *measurement perspective* (Dorans 2012; Holland 1994), the focus is on controlling potential sources of random error and potential sources of bias (e.g., construct-irrelevant score variance, construct underrepresentation, method effects). However, in any applied context, additional considerations are relevant. For example, test takers often view testing programs as contests in which they are competing for some desired outcome, and whether they achieve their goal or not, they want the process to be fair; Holland (1994) and Dorans (2012) referred to this alternate, and legitimate, point of view as the *contest perspective*.

A *pragmatic perspective* (Kane 2013b) focuses on how well the program, as implemented, achieves its goals and avoids unintended negative effects. The pragmatic perspective is particularly salient for testing programs that serve as the bases for high-stakes decisions in public contexts. To the extent that testing programs play important roles in the public arena, their claims need to be justified. The pragmatic perspective is particularly concerned about fairness but also values objectivity (defined as the absence of subjectivity or preference) as a core concern; decision makers want testing procedures to be clearly relevant, fair, and practical. In general, it is important to evaluate how well testing programs work in practice, in the contexts in which they are operating (e.g., as the basis for decisions in employment, in academic selection, in placement, in licensure and certification). Testing programs can have strong effects on individuals and institutions, both positive and negative (Frederiksen 1984). The pragmatic perspective suggests identifying those effects and explicitly weighing them against one another in considering the value, or functional worth, of a testing program.

16.9 Concluding Remarks

ETS has been heavily involved in the development of validity theory, the creation of models for validation, and the practice of validation since the organization's creation. All of the work involved in designing and developing tests, score scales, and the materials and procedures involved in reporting and interpreting scores contributes to the soundness and plausibility of the results. Similarly, all of the research conducted on how testing programs function, on how test scores are used, and on the impact of such uses on test takers and institutions contributes to the evaluation of the functional worth of programs.

This chapter has focused on the development of validity theory, but the theory developed out of a need to evaluate testing programs in appropriate ways, and therefore it has been based on the practice of assessment. At ETS, most theoretical innovations have come out of perceived needs to solve practical problems, for which the then current theory was inadequate or unwieldy. The resulting theoretical frame-

works may be abstract and complex, but they were suggested by practical problems and were developed to improve practice.

This chapter has been organized to reflect a number of major developments in the history of validity theory and practice. The validity issues and validation models were developed during different periods, but the fact that a new issue or model appeared did not generally lead to a loss of interest in the older topics and models. The issues of fairness and bias in selection and admissions were topics of interest in the early days of ETS; their conceptualization and work on them were greatly expanded in the 1960s and 1970s, and they continue to be areas of considerable emphasis today. Although the focus has shifted and the level of attention given to different topics has varied over time, the old questions have neither died nor faded away; rather, they have evolved into more general and sophisticated analyses of the issues of meaning and values that test developers and users have been grappling with for longer than a century.

Messick shaped validity theory in the last quarter of the twentieth century; therefore this chapter on ETS's contributions has given a lot of attention to his views, which are particularly comprehensive and complex. His unified, construct-based framework assumes that "validation in essence is scientific inquiry into score meaning—nothing more, but also nothing less" (Messick 1989, p. 56) and that "judging validity in terms of whether a test does the job it is employed to do … requires evaluation of the intended or unintended social consequences of test interpretation and use" (pp. 84–85). Much of the work on validity theory at the beginning of the twenty-first century can be interpreted as attempts to build on Messick's unified, construct-based framework, making it easier to apply in a straightforward way so that tests can be interpreted and used to help achieve the goals of individuals, education, and society.

Acknowledgments The authors wish to thank Randy Bennett, Cathy Wendler, and James Carlson for comments and suggestions on earlier drafts of the chapter.

References

Abelson, R. P. (1952). Sex differences in predictability of college grades. *Educational and Psychological Measurement, 12*, 638–644. https://doi.org/10.1177/001316445201200410

Allen, C. D. (1949). *Summary of validity studies of the College Entrance Examination Board tests in current use* (Research Bulletin No. RB-49-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1949.tb00872.x

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18*, 1–14. https://doi.org/10.1007/BF02289023

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H. (1987). *Does guessing really help?* (Research Report No. RR-87-16). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00220.x

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Hillsdale: Erlbaum.

Angoff, W. H., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95–106. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Angoff, W. H., & Sharon, A. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34*, 807–816. https://doi.org/10.1177/001316447403400408

Aronson, J. E. R. (1948). *April 1948 English composition reading-reliability study* (Research Bulletin No. RB-48-10). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1948.tb00912.x

Baird, L. L. (1979). *Development of an inventory of documented accomplishments for graduate admissions* (GRE Board Research Report No. 77-03R). Princeton: Educational Testing Service.

Baird, L. L., & Knapp, J. E. (1981). *The inventory of documented accomplishments for graduate admissions: Results of a field trial study of its reliability, short-term correlates, and evaluation* (GRE Board Research Report No. 78-03R). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01253.x

Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation: Final report to the National Council on Architectural Registration* (Research Memorandum No. RM-99-02). Princeton: Educational Testing Service.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2–3), 70–91. https://doi.org/10.1080/15366367.2010.508686

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer. https://doi.org/10.1007/978-1-4020-9964-9_3

Bridgeman, B., & Cahalan, C. (2007). *Time requirements for the different item types proposed for use in the revised SAT* (Research Report No. RR-07-35). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02077.x

Bridgeman, B., & Cline, F. (2000). *Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assessment* (GRE Board Report No. 96-20P). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01830.x

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41*, 137–148. https://doi.org/10.1111/j.1745-3984.2004.tb01111.x

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37–50. https://doi.org/10.1111/j.1745-3984.1994.tb00433.x

Bridgeman, B., & McHale, F. J. (1998). Potential impact of the addition of a writing assessment on admissions decisions. *Research in Higher Education, 39*, 663–677. https://doi.org/10.1023/A:1018709924672

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. S. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report No. 2000-01). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01824.x

Bridgeman, B., Cline, F., & Hessinger, J. (2004a). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education, 17*, 25–37. https://doi.org/10.1207/s15324818ame1701_2

Bridgeman, B., Trapani, C., & Curley, E. (2004b). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement, 41*, 291–310. https://doi.org/10.1111/j.1745-3984.2004.tb01167.x

Bridgeman, B., Burton, N., & Cline, F. (2008a). *Understanding what the numbers mean: A straightforward approach to GRE predictive validity* (GRE Board Research Report No. 04-03). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02132.x

Bridgeman, B., Burton, N., & Pollack, J. (2008b). Predicting grades in college courses: A comparison of multiple regression and percent succeeding approaches. *Journal of College Admission, 199*, 19–25.

Bridgeman, B., Pollack, J., & Burton, N. (2008c). *Predicting grades in different types of college courses* (Research Report No. RR-08-06). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02092.x

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33*, 267–334. https://doi.org/10.1007/BF02289327

Burton, E., & Burton, N. (1993). The effect of item screening on test scores and test characteristics. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321–336). Hillsdale: Erlbaum.

Burton, N., & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (Research Report No. RR-05-03). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01980.x

Cahalan, C., King, T. C., Cline, F., & Bridgeman, B. (2006). *Observational timing study on the SAT Reasoning Test for test-takers with learning disabilities and/or ADHD* (Research Report No. RR-06-23). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02029.x

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*, 546–553. https://doi.org/10.1037/h0048255

Campbell, J. (1964). *Testing of culturally different groups* (Research Bulletin No. RB-64-34). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1964.tb00506.x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. https://doi.org/10.1037/h0046016

Cardall, C., & Coffman, W. (1964). *A method for comparing the performance of different groups on the items in a test* (Research Bulletin No. RB-64-61). Princeton: Educational Testing Service.

Carlson, S. B., Ward, W. C., & Frederiksen, N. O. (1984). The place of social intelligence in a taxonomy of cognitive abilities. *Intelligence, 8*, 315–337. https://doi.org/10.1016/0160-2896(84)90015-1

Carroll, J. B. (1974). *Psychometric tests as cognitive tasks: A new structure of intellect* (Research Bulletin No. RB-74-16). Princeton: Educational Testing Service.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*, 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Cho, Y., & Bridgeman, B. (2012). Relationship of *TOEFL iBT* Scores to academic performance: Some evidence from American universities. *Language Testing, 29*, 421–442. https://doi.org/10.1177/0265532211430368

Cleary, A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.

Cleary, A., & Hilton, T. (1966). *An investigation of item bias* (Research Bulletin No. RB-66-17). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1966.tb00355.x

Cole, N. (1973). Bias in selection. *Journal of Educational Measurement, 5*, 237–255. https://doi.org/10.1111/j.1745-3984.1973.tb00802.x

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). New York: Macmillan.

College Board. (1948). *Report on the study adaptability ratings of cadets in the class of 1947 at the U.S. Coast Guard Academy* (Research Bulletin No. RB-48-01A). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1948.tb00001.x

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. https://doi.org/10.1037/h0040957

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Damarin, F., & Messick, S. (1965). *Response styles and personality variables: A theoretical integration of multivariate research* (Research Bulletin No. RB-65-10). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1965.tb00967.x

Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Board.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel–Haenszel method. *Applied Measurement in Education, 2*, 217–233. https://doi.org/10.1207/s15324818ame0203_3

Dorans, N. J. (2004). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37. https://doi.org/10.1111/j.1745-3992.2012.00250.x

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale: Erlbaum. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dwyer, C. A., Gallagher, A., Levin, J., & Morley, M. E. (2003). *What is quantitative reasoning? Defining the construct for assessment purposes* (Research Report No. RR-03-30). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01922.x

Ebel, R. (1962). Content standard test scores. *Educational and Psychological Measurement, 22*, 15–25. https://doi.org/10.1177/001316446202200103

Educational Testing Service. (1948). *The study of reading reliability of the College Board English composition tests of April and June 1947* (Research Bulletin No. RB-48-07). Princeton: Author. https://doi.org/10.1002/j.2333-8504.1948.tb00005.x

Educational Testing Service. (1992). *The origins of Educational Testing Service*. Princeton: Author.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs, 79*(2), 3–84.

Evans, F. R. (1980). *A study of the relationship among speed and power, aptitude test scores, and ethnic identity* (Research Report No. RR-80-22). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1980.tb01219.x

Fairtest. (2003). *SAT I: A faulty instrument for predicting college success.* Retrieved from http://fairtest.org/facts/satvalidity.html

Fan, C. T., Lord, F. M., & Tucker, L. R. (1950). *A score checking machine* (Research Memorandum No. RM-50-08). Princeton: Educational Testing Service.

Frederiksen, N. O. (1948). *The prediction of first term grades at Hamilton College* (Research Bulletin No. RB-48-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1948.tb00867.x

Frederiksen, N. (1951). The influence of timing and instructions on Cooperative Reading Test scores. *Educational and Psychological Measurement, 12*, 598–607. https://doi.org/10.1002/j.2333-8504.1951.tb00022.x

Frederiksen, N. (1959). *Development of the test "Formulating Hypotheses": A progress report* (Office of Naval Research Technical Report No. NR-2338[00]). Princeton: Educational Testing Service.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202. https://doi.org/10.1037/0003-066X.39.3.193

Frederiksen, N. (1986). Toward a broader conception of human intelligence. *American Psychologist, 41*, 445–452. https://doi.org/10.1037/0003-066X.41.4.445

Frederiksen, N., & Ward, W. (1978). Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement, 2*, 1–24. https://doi.org/10.1177/014662167800200101

Frederiksen, N., Olsen, M., & Schrader, W. (1950a). *Prediction of first-semester grades at Kenyon College, 1948–1949* (Research Bulletin No. RB-50-49). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1950.tb00879.x

Frederiksen, N., Schrader, W., Olsen, M., & Wicoff, E. (1950b). *Prediction of freshman grades at the University of Rochester* (Research Bulletin No. RB-50-44). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00481.x

Frederiksen, N., Saunders, D. R., & Wand, B. (1957). The in-basket test. *Psychological Monographs, 71*(9), 1–28. https://doi.org/10.1037/h0093706

Freeman, P. M. (1952). *Survey of studies on correction for guessing and guessing instructions* (Research Memorandum No. RM-52-04). Princeton: Educational Testing Service.

French, J. W. (1948). *Report on the study intercorrelations of entrance tests and grades, class of 1947 at the U.S. Coast Guard Academy* (Research Bulletin No. RB-48-03). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1948.tb00002.x

French, J. W. (1951a). *Conference on factorial studies of aptitude and personality measures* (Research Memorandum No. RM-51-20). Princeton: Educational Testing Service.

French, J. W. (1951b). *The description of aptitude and achievement factors in terms of rotated factors*. (Psychometric Monograph No. 5). Richmond: Psychometric Society.

French, J. W. (1954). *Manual for Kit of Selected Tests for Reference Aptitude and Achievement Factors*. Princeton: Educational Testing Service.

French, J. W. (1957). *The relation of ratings and experience variables to Tests of Developed Abilities profiles* (Research Bulletin No. RB-57-14). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1957.tb00934.x

French, J. W. (1962). Effect of anxiety on verbal and mathematical examination scores. *Educational and Psychological Measurement, 22*, 553–564. https://doi.org/10.1177/001316446202200313

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual for Kit of Reference Tests for Cognitive Factors*. Princeton: Educational Testing Service.

Green, B. F., Jr. (1950). A note on the calculation of weights for maximum battery reliability. *Psychometrika, 15*, 57–61. https://doi.org/10.1007/BF02289178

Green, B. F., Jr. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika, 17*(4), 429–440. https://doi.org/10.1007/BF02288918

Green, D. R. (1982). Methods used by test publishers to "debias" standardized tests. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 278–313). Baltimore: The Johns Hopkins University Press.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Gulliksen, H. (1950a). *Intrinsic versus correlational validity* (Research Bulletin No. RB-50-37). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1950.tb00477.x

Gulliksen, H. (1950b). *Theory of mental tests*. New York: Wiley. https://doi.org/10.1037/13240-000

Gulliksen, H., & Messick, S. (1960). *Psychological scaling: Theory and applications*. New York: Wiley.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229. https://doi.org/10.3102/1076998607302636

Harman, H. H. (1967). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Holland, P. W. (1994). Measurements or contests? Comment on Zwick, Bond, and Allen/Donogue. In *Proceedings of the Social Statistics Section of the American Statistical Association* (pp. 27–29). Alexandria: American Statistical Association.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_2

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport: American Council on Education and Praeger.
Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating*. New York: Academic Press.
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale: Erlbaum.
Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale: Erlbaum.
Horst, P. (1950a). *Optimal test length for maximum battery validity* (Research Bulletin No. RB-50-36). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1950.tb00476.x
Horst, P. (1950b). *The relationship between the validity of a single test and its contribution to the predictive efficiency of a test battery* (Research Bulletin No. RB-50-32). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1950.tb00472.x
Horst, P. (1951). Estimating total test reliability from parts of unequal length. *Educational and Psychological Measurement, 11*, 368–371. https://doi.org/10.1177/001316445101100306
Jackson, D. (2002). The constructs in people's heads. In H. I. Braun, D. N. Jackson, D. E. Wiley, & S. Messick (Eds.), *The role of constructs in psychological and educational measurement* (pp. 3–17). Mahwah: Erlbaum.
Jackson, D., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*, 243–252. https://doi.org/10.1037/h0045996
Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement, 21*, 771–790. https://doi.org/10.1177/001316446102100402
Jackson, D. N., & Messick, S. (1965). The person, the product, and the response: Conceptual problems in the assessment of creativity. *Journal of Personality, 33*, 309–329. https://doi.org/10.1111/j.1467-6494.1965.tb01389.x
Jackson, D., Messick, S., & Solley, C. (1957). A multidimensional scaling approach to the perception of personality. *Journal of Psychology, 22*, 311–318. https://doi.org/10.1080/00223980.1957.9713088
Johnson, A. P., & Olsen, M. A. (1952). *Comparative three-year and one-year validities of the Law School Admissions Test at two law schools* (Law School Admissions Council Report No. 52-02). Princeton: Educational Testing Service.
Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482. https://doi.org/10.1007/BF02289658
Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202. https://doi.org/10.1007/BF02289343
Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.
Jöreskog, K. G., & Lawley, D. N. (1967). *New methods in maximum likelihood factor analysis* (Research Bulletin No. RB-67-49). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1967.tb00703.x
Jöreskog, K. G., & van Thillo, M. (1972). *LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (Research Bulletin No. RB-72-56). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1972.tb00827.x
Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527
Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger.
Kane, M. (2012). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*, 3–17. https://doi.org/10.1177/0265532211417210
Kane, M. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. (2013b). Validity and fairness in the testing of individuals. In M. Chatterji (Ed.), *Validity and test use* (pp. 17–53). Bingley: Emerald.
Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22*, 29–41. https://doi.org/10.1007/BF02289207
Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers: World Book.
Klieger, D. M., Holtzman, S., & Ezzo, C. (2013, April). *The promise of non-cognitive assessment in graduate and professional school admissions*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
Kristof, W. (1962). *Statistical inferences about the error variance* (Research Bulletin No. RB-62-21). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1962.tb00299.x
Kristof, W. (1970). On the sampling theory of reliability estimation. *Journal of Mathematical Psychology, 7*, 371–377. https://doi.org/10.1016/0022-2496(70)90054-4
Kristof, W. (1971). On the theory of a set of tests which differ only in length. *Psychometrika, 36*, 207–225. https://doi.org/10.1007/BF02297843
Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39*, 491–499.
Kyllonen, P. (2005, September). The case for noncognitive assessments. *ETS R&D Connections*, pp. 1–7.
Lawrence, I., & Shea, E. (2011, April). *A brief history of Educational Testing Service as a scientific organization*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.
Lemann, N. (1995). The great sorting. *Atlantic Monthly, 276*(3), 84–100.
Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, DC: American Council on Education.
Linn, R. L. (1972). *Some implications of the Griggs decision for test makers and users* (Research Memorandum No. RM-72-13). Princeton: Educational Testing Service.
Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161. https://doi.org/10.3102/00346543043002139
Linn, R. L. (1975). *Test bias and the prediction of grades in law school* (Research Report No. 75-01). Newtown: Law School Admissions Council.
Linn, R. L. (1976). In search of fair selection procedures. *Journal of Educational Measurement, 13*, 53–58. https://doi.org/10.1111/j.1745-3984.1976.tb00181.x
Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21*, 33–47. https://doi.org/10.1111/j.1745-3984.1984.tb00219.x
Linn, R. L., & Werts, C. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1–4. https://doi.org/10.1111/j.1745-3984.1971.tb00898.x
Liu, J., Allsbach, J. R., Feigenbaum, M., Oh, H.-J., & Burton, N. W. (2004). *A study of fatigue effects from the New SAT* (Research Report No. RR-04-46). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01973.x
Livingston, S. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9*, 13–26. https://doi.org/10.1111/j.1745-3984.1972.tb00756.x
Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635–694.
Lord, F. M. (1951). *A theory of test scores and their relation to the trait measured* (Research Bulletin No. RB-51-13). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1951.tb00922.x
Lord, F. M. (1953). *Speeded tests and power tests—An empirical study of validities* (Research Bulletin No. RB-53-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1953.tb00228.x

Lord, F. M. (1955). *Estimating test reliability* (Research Bulletin No. RB-55-07). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1955.tb00054.x

Lord, F. M. (1956). *Do tests of the same length have the same standard error of measurement?* (Research Bulletin No. RB-56-07). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1956.tb00063.x

Lord, F. M. (1957). *Inferring the shape of the frequency distribution of true scores and of errors of measurement* (Research Bulletin No. RB-57-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1957.tb00076.x

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement, 19*, 233–239. https://doi.org/10.1177/001316445901900208

Lord, F. M. (1961). *Estimating norms by item sampling* (Research Bulletin No. RB-61-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1961.tb00103.x

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika, 30*, 239–270. https://doi.org/10.1007/BF02289490

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–305. https://doi.org/10.1037/h0025105

Lord, F. M. (1974). *Formula scoring and number-right scoring* (Research Memorandum No. RM-52-04). Princeton: Educational Testing Service.

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

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

Lord, F. M., & Stocking, M. (1976). An interval estimate for making statistical inferences about true score. *Psychometrika, 41*, 79–87. https://doi.org/10.1007/BF02291699

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 453–461. https://doi.org/10.1177/014662168400800409

Markle, R., Olivera-Aguilar, M., Jackson, R., Noeth, R., & Robbins, S. (2013). *Examining evidence of reliability, validity, and fairness for the Success Navigator Assessment* (Research Report No. RR-13-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02319.x

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of the Advanced Placement Examinations* (College Board Report No. 92-07). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01516.x

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681–689). Westport: American Council on Education and Praeger.

Messick, S. (1956). The perception of social attitudes. *Journal of Abnormal and Social Psychology, 52*, 57–66. https://doi.org/10.1037/h0038586

Messick, S. (1962). Response style and content measures from personality inventories. *Educational and Psychological Measurement, 22*, 41–56. https://doi.org/10.1177/001316446202200106

Messick, S. (1964). Personality measurement and college performance. In A. G. Wesman (Ed.), *Proceedings of the 1963 invitational conference on testing problems* (pp. 110–129). Princeton: Educational Testing Service.

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist, 20*, 136–142. https://doi.org/10.1037/h0021712

Messick, S. (1967). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 115–145). Chicago: Aldine.

Messick, S. (1970). The criterion problem in the evaluation of instruction: Assessing possible, not just intended outcomes. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction: Issues and problems* (pp. 183–202). New York: Holt, Rinehart and Winston.

Messick, S. (1972). Beyond structure: In search of functional models of psychological process. *Psychometrika, 37*, 357–375. https://doi.org/10.1007/BF02291215

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966. https://doi.org/10.1037/0003-066X.30.10.955
Messick, S. (1977). Values and purposes in the uses of tests. In *Speaking out: The use of tests in the policy arena* (pp. 20–26). Princeton: Educational Testing Service.
Messick, S. (1979). *Test validity and the ethics of assessment*. (Research Report No. RR-79-10). Princeton: Educational Testing Service.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012
Messick, S. (1981a). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*, 575–588. https://doi.org/10.1037/0033-2909.89.3.575
Messick, S. (1981b). The controversy over coaching: Issues of effectiveness and equity. *New Directions for Testing and Measurement, 11*, 21–53.
Messick, S. (1982a). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing policy. *Educational Psychologist, 17*, 69–91. https://doi.org/10.1080/00461528209529246
Messick, S. (1982b). The values of ability testing: Implications of multiple perspectives about criteria and standards. *Educational Measurement: Issues and Practice, 1*, 9–12. https://doi.org/10.1111/j.1745-3992.1982.tb00660.x
Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist, 19*, 59–74. https://doi.org/10.1080/00461528409529283
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Erlbaum.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science—A volume in honor of Lee J. Cronbach* (pp. 161–200). Hillsdale: Erlbaum.
Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13–23. https://doi.org/10.3102/0013189X023002013
Messick, S. (1994b). *Standards-based score interpretation: Establishing valid grounds for valid inferences* (Research Report No. RR-94-57). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01630.x
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x
Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35–44. https://doi.org/10.1023/A:1006964925094
Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 3–20). Boston: Kluwer. https://doi.org/10.1007/978-1-4615-4397-8_1
Messick, S., & Abelson, R. (1957). Research tools: Scaling and measurement theory. *Review of Educational Research, 27*, 487–497. https://doi.org/10.2307/1169167
Messick, S., & Anderson, S. (1970). Educational testing, individual development, and social responsibility. *The Counseling Psychologist, 2*, 80–88. https://doi.org/10.1177/001100007000200215
Messick, S., & Frederiksen, N. (1958). Ability, acquiescence, and "authoritarianism." *Psychological Reports, 4*, 687–697. https://doi.org/10.2466/pr0.1958.4.3.687
Messick, S., & Jackson, D. (1958). The measurement of authoritarian attitudes. *Educational and Psychological Measurement, 18*, 241–253. https://doi.org/10.1177/001316445801800202
Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216. https://doi.org/10.1037/0033-2909.89.2.191

Messick, S., & Kogan, N. (1963). Differentiation and compartmentalization in object-sorting measures of categorizing style. *Perceptual and Motor Skills, 16*, 47–51. https://doi.org/10.2466/pms.1963.16.1.47

Messick, S., & Ross, J. (Eds.). (1962). *Measurement in personality and cognition*. New York: Wiley.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report No. 83-1). Princeton: Educational Testing Service.

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

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483. https://doi.org/10.1007/BF02294388

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416. https://doi.org/10.1111/j.1745-3984.1996.tb00498.x

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–306). Westport: American Council on Education and Praeger.

Mislevy, R. (2007). Validity by design. *Educational Researcher, 36*, 463–469. https://doi.org/10.3102/0013189X07311660

Mislevy, R. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 83–108). Charlotte: Information Age.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Princeton: Educational Testing Service.

Mislevy, R., Steinberg, L., & Almond, R. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97–128). Mahwah: Erlbaum.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003a). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003b). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R., Haertel, G., Cheng, B., Ructtinger, L., DeBarger, A., Murray, E., et al. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation, 19*, 121–140. https://doi.org/10.1080/13803611.2013.767614

Mollenkopf, W. G. (1950). Predicted differences and differences between predictions. *Psychometrika, 15*, 259–269. https://doi.org/10.1002/j.2333-8504.1950.tb00018.x

Mollenkopf, W. G. (1951). *Effectiveness of Naval Academy departmental standings for predicting academic success at the U.S. Naval Postgraduate School* (Research Bulletin No. RB-51-22). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1951.tb00221.x

Myers, A. E. (1965). Risk taking and academic success and their relation to an objective measure of achievement motivation. *Educational and Psychological Measurement, 25*, 355–363. https://doi.org/10.1177/001316446502500206

Myers, C. T. (1975). *Test fairness: A comment on fairness in statistical analysis* (Research Bulletin No. RB-75-12). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1975.tb01051.x

Novick, M. R. (1965). *The axioms and principal results of classical test theory* (Research Report No. RR-65-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1965.tb00132.x

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13. https://doi.org/10.1007/BF02289400

Novick, M. R., & Thayer, D. T. (1969). *Some applications of procedures for allocating testing time* (Research Bulletin No. RB-69-01). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00161.x

Packer, J. (1987). *SAT testing time for students with disabilities* (Research Report No. RR-87-37). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00241.x

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_4

Petersen, N., & Novick, M. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29. https://doi.org/10.1111/j.1745-3984.1976.tb00178.x

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Popper, K. (1965). *Conjectures and refutations*. New York: Harper and Row.

Porter, T. (2003). Measurement, objectivity, and trust. *Measurement: Interdisciplinary Research and Perspectives, 1*, 241–255. https://doi.org/10.1207/S15366359MEA0104_1

Powers, D. E. (1988). Incidence, correlates, and possible causes of test anxiety in graduate admissions testing. *Advances in Personality Assessment, 7*, 49–75.

Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the *Graduate Record Examinations*® (GRE) General Test. *Journal of Educational Computing Research, 24*, 249–273. https://doi.org/10.2190/680W-66CR-QRP7-CL1F

Ragosta, M., & Wendler, C. (1992). *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees* (Research Report No. RR-92-35). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01466.x

Ramist, L., & Weiss, G. E. (1990). The predictive validity of the SAT, 1964 to 1988. In W. W. Willingham, C. Lewis, R. L. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 117–140). Princeton: Educational Testing Service.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (Research Report No. RR-94-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01600.x

Rikoon, S., Liebtag, T., Olivera-Aguilar, M., Robbins, S., & Jackson, T. (2014). *A pilot study of holistic assessment and course placement in community college samples: Findings and recommendations* (Research Memorandum No. RM-14-10). Princeton: Educational Testing Service.

Rivera, C., & Schmitt, A. P. (1988). *A comparison of Hispanic and White non-Hispanic students' omit patterns on the Scholastic Aptitude Test* (Research Report No. RR-88-44). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00300.x

Roberts, R. D., Schulze, R., & MacCann, C. (2008). The measurement of emotional intelligence: A decade of progress? In G. Boyle, G. Matthews, & D. Saklofske (Eds.), *The Sage handbook of personality theory and assessment: Vol. 2. Personality measurement and testing* (pp. 461–482). London: Sage. https://doi.org/10.4135/9781849200479.n22

Ryans, D. G., & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 455–494). Washington, DC: American Council on Education.

Schiffman, H., & Messick, S. (1963). Scaling and measurement theory. *Review of Educational Research, 33*, 533–542. https://doi.org/10.2307/1169654

Schrader, W. B. (1965). A taxonomy of expectancy tables. *Journal of Educational Measurement, 2*, 29–35. https://doi.org/10.1111/j.1745-3984.1965.tb00388.x

Schultz, D. G. (1952). Item validity and response change under two different testing conditions. *Journal of Educational Psychology, 45*, 36–43. https://doi.org/10.1037/h0059845

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405–450). Washington, DC: American Educational Research Association. https://doi.org/10.2307/1167347

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36*, 1138–1146. https://doi.org/10.1037/0003-066X.36.10.1138

Shimberg, B. (1982). *Occupational licensing: A public perspective* (Center for Occupational and Professional Assessment Report). Princeton: Educational Testing Service.

Shimberg, B. (1990). Social considerations in the validation of licensing and certification exams. *Educational Measurement: Issues and Practice, 9*, 11–14. https://doi.org/10.1111/j.1745-3992.1990.tb00386.x

Stricker, L. J. (2008). *The challenge of stereotype threat for the testing community* (Research Memorandum No. RM-08-12). Princeton: Educational Testing Service.

Stricker, L. J., & Bejar, I. (2004). Test difficulty and stereotype threat on the GRE General Test. *Journal of Applied Social Psychology, 34*, 563–597. https://doi.org/10.1111/j.1559-1816.2004.tb02561.x

Stricker, L. J., & Rock, D. A. (1990). Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences, 11*, 833–839. https://doi.org/10.1016/0191-8869(90)90193-U

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665–693. https://doi.org/10.1111/j.1559-1816.2004.tb02564.x

Stricker, L., Messick, S., & Jackson, D. (1969). *Conformity, anticonformity, and independence: Their dimensionality and generality* (Research Bulletin No. RB-69-17). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb01006.x

Stricker, L. J., Rock, D. A., & Bennett, R. E. (2001). Sex and ethnic-group differences on accomplishments measures. *Applied Measurement in Education, 14*, 205–218. https://doi.org/10.1207/S15324818AME1403_1

Tucker, L. R. (1948). *Memorandum concerning study of effects of fatigue on afternoon achievement test scores due to scholastic aptitude test being taken in the morning* (Research Memorandum No. RM-48-02). Princeton: Educational Testing Service.

Tucker, L. R. (1949). A note on the estimation of test reliability by the Kuder–Richardson formula (20). *Psychometrika, 14*, 117–119. https://doi.org/10.1007/BF02289147

Tucker, L. R. (1955). The objective definition of simple structure in linear factor analysis. *Psychometrika, 20*, 209–225. https://doi.org/10.1007/BF02289018

Turnbull, W. (1949). Influence of cultural background on predictive test scores. In *Proceedings of the ETS invitational conference on testing problems* (pp. 29–34). Princeton: Educational Testing Service.

Turnbull, W. (1951). Socio-economic status and predictive test scores. *Canadian Journal of Psychology, 5*, 145–149. https://doi.org/10.1037/h0083546

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking*. New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer. https://doi.org/10.1007/b97446

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

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

Wigdor, A., & Garner, W. (1982). *Ability testing: Uses, consequences, and controversies*. Washington, DC: National Academy Press.

Wild, C., & Dwyer, C. A. (1980). Sex bias in selection. In L. van der Kamp, W. Langerat, & D. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 153–168). New York: Wiley.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. New York: College Entrance Examination Board.

Willingham, W. W. (1999). A systematic view of test fairness. In S. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 213–242). Mahwah: Erlbaum.

Willingham, W. W., & Breland, H. M. (1982). *Personal qualities and college admissions*. New York: College Entrance Examination Board.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah: Erlbaum.

Willingham, W. W., Bennett, R. E., Braun, H. I., Rock, D. A., & Powers, D. A. (Eds.). (1988). *Testing handicapped people*. Needham Heights: Allyn and Bacon.

Willingham, W., Lewis, C., Morgan, R., & Ramist, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades*. Princeton: Educational Testing Service.

Wohlhueter, J. F. (1966). *Fatigue in testing and other mental tasks: A literature survey* (Research Memorandum No. RM-66-06). Princeton: Educational Testing Service.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*, 147–170. https://doi.org/10.1177/0265532209349465

Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.), *Rethinking the SAT* (pp. 289–301). New York: Routledge Falmer.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Mahwah: Erlbaum.

Zieky, M. (2006). Fairness review. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Hillsdale: Erlbaum.

Zieky, M. (2011). The origins of procedures for using differential item functioning statistics at Educational Testing Service. In N. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul Holland* (pp. 115–127). New York: Springer. https://doi.org/10.1007/978-1-4419-9389-2_7

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647–679). Westport: American Council on Education and Praeger.

Chapter 17
Understanding the Impact of Special Preparation for Admissions Tests

Donald E. Powers

By examining unique developments and singular advancements, it is possible to sort the history of educational and psychological testing into a number of distinct phases. One topic that seems to permeate all stages, however, is the question of how best to *prepare* for such tests. This chapter documents some of Educational Testing Service's (ETS's) contributions to understanding the role of test preparation in the testing process. These contributions include (a) analyzing key features of test preparation, (b) understanding the effects of various sorts of preparation on test performance, and (c) devising tests that will yield meaningful scores in the face of both legitimate as well as questionable attempts to improve test-taker performance. The chapter begins with a definition of special test preparation and then elaborates on its significance. Next, it examines the nature of interest in the topic. Finally, it explores ETS Research and Development (R&D) contributions to explicating the issues associated with special test preparation.

17.1 Definitions

The first issue that one encounters when discussing test preparation is terminology. This terminology applies both to the tests that are involved and to the kinds of preparation that are directed at test takers. Most of the research described below pertains to several tests that are designed to measure academic abilities (e.g., verbal and quantitative reasoning abilities) that develop relatively slowly over a significant

This chapter was originally published in 2012 by Educational Testing Service as a research report in the ETS R&D Scientific and Policy Contributions Series.

D.E. Powers (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: dpowers@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_17

period of time. This improvement occurs as a result of both formal schooling as well as other less formal experiences outside of school. Thus, to varying degrees, all students who take these kinds of tests receive highly relevant (but certainly differentially effective) preparation that should improve the skills and abilities being tested.

With respect to preparation, we have chosen here to use the word *special* to refer to a particular category of test preparation that focuses on readying test takers for a specific test. This special preparation may be of different sorts. For example, test familiarization is designed to ensure that prospective test takers are well versed in the general skills required for test taking and to help them gain familiarity with the procedures that are required to take a particular test. This type of preparation may entail, for instance, exposing test takers to the kinds of item formats they will encounter, making certain that they know when to guess, and helping them learn to apportion their time appropriately. Special preparation of this sort is generally regarded as desirable, as it presumably enables individuals to master the mechanics of test taking, thereby freeing them to focus on, and accurately demonstrate, the skills and abilities that are being assessed.

Coaching, on the other hand, has had a decidedly more negative connotation insofar as it is typically associated with short-term efforts aimed at teaching test-taking strategies or "tricks" to enable test takers to "beat the test;" that is, to take advantage of flaws in the test or in the testing system (e.g., never choose a particular answer choice if a question has these characteristics…). As Messick (1982) has noted, however, the term *coaching* has often been used in a variety of ways. At one extreme, it may signify short-term cramming and practice on sample item types, while on the other it may denote long-term instruction designed to develop the skills and abilities that are being tested. In practice, the distinctions among (a) relevant instruction, (b) test familiarization, and (c) coaching are sometimes fuzzy, as many programs contain elements of each type of preparation.

17.1.1 Significance of Special Test Preparation

Messick (1982) noted three ways in which special preparation may improve test scores. Each of these ways has a very different implication for score use. First, like real instruction, some types of special test preparation may genuinely improve the skills and abilities being tested, thereby resulting in higher test scores also. This outcome should have no detrimental effect on the validity of scores.

Second, some special test preparation (or familiarization) may enhance general test-taking skills and reduce test anxiety, thereby increasing test scores that may otherwise have been inaccurately low indicators of test takers' true abilities. Insofar as this kind of preparation reduces or eliminates unwanted sources of test difficulty, it should serve only to improve score validity.

The third possibility is that if it entails the teaching of test-taking tricks or other such strategies, special test preparation may increase test scores without necessarily

improving the underlying abilities that are being assessed. A likely result is inaccurately high test scores and diminished score validity.

Finally, along with score validity, equity is often at issue in special test preparation, as typically not all students have equal opportunity to benefit in the ways described above. If special preparation is effective, its benefits may accrue only to those who can afford it.

17.1.2 Interest in Special Test Preparation

At first blush, the issue of special test preparation might seem to be of interest mainly to a relatively small group of test developers and psychometricians. Historically, however, attention to this topic has been considerably more widespread. Naturally, test takers (and for some tests, their parents) are concerned with ensuring that they are well prepared to take any tests that have high-stakes consequences. However, other identifiable groups have also shown considerable interest in the topic.

For instance, concern is clearly evident in the professional community. The current version of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014) suggests a need to establish the degree to which a test is susceptible to improvement from special test preparation (Standard 1.7: "If test performance, or a decision made therefrom, is claimed to be essentially unaffected by practice and coaching, then the propensity for test performance to change with these forms of instruction should be documented," p. 24). In addition, a previous edition of *Educational Measurement* (Linn 1989), perhaps the most authoritative work on educational testing, devoted an entire chapter to special test preparation (Bond 1989).

General public interest is apparent also, as coaching has been the subject of numerous articles in the popular media (e.g., "ETS and the Coaching Cover Up," Levy 1979). One study of the effects of coaching (Powers and Rock 1999) was even a topic of discussion on a prominent national television show when the host of the *Today Show*, Matt Lauer, interviewed College Board Vice President Wayne Camara.

Besides being of general interest to the public, ETS coaching studies have also had a major impact on testing policy and practice. For example, in the early 1980s a previously offered section of the *GRE*® General Test (the analytical ability measure) was changed radically on the basis of the results of a GRE Board-sponsored test preparation study (Powers and Swinton 1984).

As a final indication of the widespread interest in the topic, in the late 1970s the U.S. Federal Trade Commission (FTC) became so troubled by the possibly misleading advertising of commercial coaching companies that it launched a major national investigation of the efficacy of such programs (Federal Trade Commission 1978, 1979). As described below, ETS contributed in several ways to this effort.

17.2 Studying the Effects of Special Test Preparation

What follows is an account of several key ETS contributions to understanding the role and effects of special test preparation. The account is organized within each of the two major testing programs on which special test preparation research has concentrated, the *SAT*® test and the GRE General Test.

17.2.1 The SAT

17.2.1.1 The College Board Position

The *effectiveness* of special test preparation has long been a contentious issue. Perhaps a reasonable place to begin the discussion is with the publication of the College Board's stance on coaching, as proclaimed by the Board's trustees in *Effects of Coaching on Scholastic Aptitude Test Scores* (College Entrance Examination Board 1965). This booklet summarized the (then) relatively few, mostly ETS-sponsored studies of coaching for the SAT (e.g., Dyer 1953, and French and Dear 1959) and concluded, "... the magnitude of gains resulting from coaching vary slightly, but they are always small ..." (p. 4), the average gain being fewer than 10 points on the 200-800 SAT scale.

17.2.1.2 Early Studies

The first significant challenge to the Board's stance seems to have come with the completion of a study by ETS researchers Evans and Pike (1973), who demonstrated that two SAT quantitative item types being considered for inclusion in the SAT were susceptible to improvement through special preparation—in particular, to the Saturday morning test preparation classes that the researchers designed for implementation over a 7-week period. The researchers' best estimate of effects was about 25 points on the 200–800 SAT Math (SAT-M) scale.

Besides the significant program of instruction that Evans and Pike developed, another particularly noteworthy aspect of this effort was the researchers' ability to implement a true experimental design. Students were randomly assigned to either (a) one of three treatment groups, each of which focused specifically on a different item type, or (b) a comparison condition that involved only more general test-taking skills. Previously, virtually no such studies had successfully carried out a true experiment.

At least partly because of the Evans and Pike (1973) study, interest also increased in the effects of special preparation for the verbal section of the SAT. The College Board subsequently funded ETS researchers to study the effectiveness of special secondary school programs geared to improving SAT Verbal (SAT-V) scores (Alderman and Powers 1980). A contribution here was that instead of relying on

strictly observational methods or quasi-experimental designs, the investigators were able, through careful collaboration with a set of secondary schools, to exert a reasonably strong degree of experimental control over *existing* special preparation programs, assigning students randomly to treatment or control groups. This task was accomplished, for example, by taking advantage of demand for preparation that, in some cases, exceeded the schools' ability to offer it. In other cases, it was possible to simply delay preparation for randomly selected students. The results suggested that secondary school programs can affect SAT-V scores, albeit modestly, increasing them by about 4–16 points on the 200–800 SAT-V scale.

17.2.1.3 Test Familiarization

About the same time, the College Board, realizing the need to ensure that all test takers were familiar with the SAT, developed a much more extensive information bulletin than had been available previously. The new booklet, called *Taking the SAT*, contained extensive information about the test and about test-taking strategies, a review of math concepts, and a full-length practice SAT. Much to its credit, the Board was interested not only in offering the more extensive preparation material, but also in learning about its impact, and so it commissioned a study to assess the booklet's effects on both test-taking behavior and test scores (Powers and Alderman 1983). The study was an experiment in which a randomly selected group of SAT registrants received a prepublication version of the new booklet. Subsequently, their test performance was compared with that of an equivalent randomly selected group of test takers who had not received the booklet. (Only high school juniors were included in the study, partly to ensure that, should the booklet prove effective in increasing scores, all students in the cohort would have the opportunity to benefit from it before they graduated.)

The results showed increases in knowledge of appropriate test-taking behavior (e.g., when to guess), decreased anxiety, and increased confidence. There were no statistically significant effects on SAT-V scores but a small, significant effect on SAT-M scores of about 8 points.

17.2.1.4 Federal Interest

Perhaps the single most significant factor in the rising interest in coaching and test preparation was the involvement of the U.S. Federal Trade Commission (FTC). The FTC became increasingly concerned about the veracity of claims being made by commercial coaching companies, which promised to increase SAT takers' scores by hundreds of points. The issue became so important that the FTC eventually undertook its own study to investigate the effectiveness of commercial coaching programs.

Both ETS and several of the major commercial coaching companies cooperated with the FTC investigation. ETS provided students' SAT scores, and the coaching

companies provided information about students' enrollment in their programs. FTC researchers analyzed the data and eventually issued a report, finding the effects of commercial coaching for the SAT to be statistically significant—in the range of 20–30 points for both SAT-V and SAT-M at the most effective of the coaching schools that were studied (Federal Trade Commission 1978, 1979; Sesnowitz et al. 1982). Needless to say, the study attracted considerable attention.

ETS responded to the FTC's findings as follows. Samuel Messick, then Vice President for Research at ETS, assembled a team of researchers to take a critical look at the methods the FTC had used and the conclusions it had reached. Messick and his team critiqued the FTC's methodology and, in order to address some serious flaws in the FTC analyses, reanalyzed the data. Various methods were employed to correct mainly for test taker self-selection in attending coaching programs.

Messick's contribution was released as a monograph titled, "The Effectiveness of Coaching for the SAT: Review and Reanalysis of Research from the Fifties to the FTC" (Messick 1980). In the book, Messick summarized and critiqued previous research on coaching, and several ETS researchers offered their critiques of the FTC study. Most importantly, the researchers conducted several reanalyses of the data obtained from the FTC. For example, ETS consultant Thomas Stroud reanalyzed the data, controlling for a variety of background variables, and found results similar to those reported by the FTC. In addition, by considering *PSAT/NMSQT*® scores, as well as pre- and postcoaching SAT scores, ETS researcher Don Rock was able to apply a differential growth model to the FTC data. His analysis showed that, at least for SAT-V scores, some of the difference between the posttest SAT scores of coached and uncoached test takers could be attributed, not to any specific effect of coaching, but rather to the faster growth expected of coached students. (The differential growth rate of coached and uncoached students was determined from PSAT/NMSQT to SAT score changes *before* students were coached.) The results of the various ETS analyses differed somewhat, but in total they revealed that only one of the three coaching schools had a significant impact on SAT scores—about 12–18 points on the SAT-V scale and about 20–30 points on the SAT-M scale.

One of the main lessons from the critique and reanalysis of the FTC study was stated by Messick (1980) in the preface to the report. Messick wrote that the issue of the effectiveness of coaching for the SAT is much more complicated than the simplistic question of whether coaching works or not. Coaching in and of itself is not automatically to be either rejected or encouraged. Rather, it matters what materials and practices are involved, at what cost in student time and resources, and with what effect on student skills, attitudes, and test scores (p. v).

Messick's (1980) insight was that complex issues, like the coaching controversy, are rarely ever usefully framed as either/or, yes/no questions. Rather, those questions turn out to involve degrees and multiple factors that need to be appreciated and sorted out. As a consequence, the answer to most questions is usually not a simple "yes" or "no," but more often a sometimes frustrating, "it depends." The task of researchers, then, is usually to determine, as best they can, the factors on which the effects depend.

17.2.1.5 Extending Lessons Learned

Messick followed through with this theme by analyzing the relationship of test preparation effects to the duration or length of test preparation programs. He published these results in the form of a meta-analysis (Messick and Jungeblut 1981), in which the authors noted "definite regularities" (p. 191) between SAT coaching effects and the amount of student contact time in coaching programs. On this basis, Messick and Jungeblut concluded that the size of the effects being claimed by coaching companies could probably be obtained only with programs that were tantamount to full-time schooling.

Powers (1986) followed Messick and Jungeblut's (1981) lead by reviewing a variety of other features of test preparation and coaching programs, and relating these features to the size of coaching effects. The advance here was that instead of focusing on the features of coaching programs, Powers analyzed the characteristics of the *item types* that comprised a variety of tests—for instance, how complex their directions were, whether they were administered under timed or untimed conditions, and what kinds of formats they employed. The results suggested that some features of test items (e.g., the complexity of directions) did render them more susceptible to improvement through coaching and practice than did others.

Several of the studies that Powers reviewed were so-called within-test practice studies, which were conducted by ETS statistical analysts (e.g., Faggen and McPeek 1981; Swinton et al. 1983; Wightman 1981). This innovative method involved trying out new test item types in early and later sections of the same test form. Then, differences in performance were compared for these early and later administered items. For some item types, it was routinely noticed that examinees performed better on new items that appeared later in the test, after earlier appearances of items of that type. A large within-test practice effect was viewed as a sufficient condition to disqualify a proposed new item type from eventual operational use. The rationale was the following: If an item type exhibited susceptibility to simple practice *within* a single test session, surely it would be *at least* as susceptible to more intensive coaching efforts.

17.2.1.6 Studying the 1994 Revision to the SAT

In 1994, a revision of the SAT was introduced. Many of the changes suggested that the revision should be even less susceptible to coaching than the earlier version. However, claims being made by coaching companies did not subside. For example, the January, 8, 1995, issue of the Philadelphia *Inquirer* proclaimed "New SAT proves more coachable than old." At least partly in response to such announcements, the College Board sponsored research to examine the effects of commercial coaching on SAT scores. Powers and Rock (1999) surveyed SAT takers about their test preparation activities, identifying a subset of test takers who had attended commercial coaching programs. Although the study was observational in nature, the researchers obtained a wide variety of background information on test takers and

used this information to control statistically for self-selection effects. This approach was necessary, as it was widely acknowledged that coached and uncoached students differ on numerous factors that are also related to SAT scores. One of the differences noted by Powers and Rock, and controlled in their analysis, was that coached test takers were more likely than their uncoached counterparts to have engaged in a variety of *other* test preparation activities (e.g., self-study of various sorts), which may also have affected SAT scores. Several alternative analyses were employed to control for self-selection effects, and although each of the analyses produced slightly different estimates, all of them suggested that the effects of coaching were far less than was being alleged by coaching enterprises—perhaps only a quarter as large as claimed.

The alternative analyses yielded coaching effect estimates of 6–12 for SAT-V and 13–26 points for SAT-M. When analyses were undertaken separately for major coaching companies, the results revealed SAT-V effects of 12–19 points for one company and 5–14 points for another. The effects for SAT-M were 5–17 and 31–38, respectively, suggesting that the two programs were differentially effective for the two portions of the SAT.

The results of the study were featured in a New York *Times* article (Bronner 1998). The article quoted Professor Betsy Jane Becker, who had reviewed numerous SAT coaching studies (Becker 1990), as saying that the study was "perhaps the finest piece of coaching research yet published" (p. A23). This assessment may of course reflect either a regard for the high quality of the study or, on the other hand, concern about the limitations of previous ones.

17.2.2 The GRE General Test

Although the SAT program has been a major focus of test preparation and coaching studies, the GRE Board has also sponsored a number of significant efforts by ETS researchers. For instance, the GRE program revised its General Test in the late 1970s, introducing an analytical ability measure to complement the long-offered verbal and quantitative reasoning measures (Powers and Swinton 1981). Concurrently, the GRE Board sponsored several studies to examine the susceptibility of the new measure to coaching and other forms of special test preparation. Swinton and Powers (1983) designed a brief course to prepare students for the new analytical section of the GRE General Test and offered it to a small group of volunteer GRE test takers at a local university. Controlling for important pre-existing differences between groups, they compared the postcourse GRE performance of these specially prepared individuals with that of all other GRE test takers at the same university. They found that the specially prepared group did much better on the analytical section (by about 66 points on the 200–800 scale) than did the larger comparison group, even after controlling for differences in the GRE verbal and quantitative scores of the two groups.

Powers and Swinton (1984) subsequently packaged the course and used it in an experimental study in which a randomly selected sample of GRE test takers received the course materials by mail. A comparison of the test scores of the prepared sample with those of a randomly selected equivalent sample of nonprepared GRE test takers revealed score improvements that were nearly as large (about 53 points with about 4 hours of self-preparation) as those observed in the face-to-face classroom preparation. A major implication of this latter study was that test preparation designed for self-study by test takers themselves was a viable alternative to more expensive, formal face-to-face interventions. The ramifications for fairness and equity were obvious. However, although the researchers were relatively sanguine about the prospects for ensuring that all examinees could be well prepared for the "coachable" item types on the GRE, the GRE Board took a conservative stance, deciding instead to remove the two most susceptible item types from the analytical ability measure.

Data collected in the studies of the GRE analytical measure were also used to gauge the effectiveness of formal commercial coaching for the verbal and quantitative sections (Powers 1985a). That is, since the analytical measure had been shown to be coachable, it could serve as a baseline against which to judge the coachability of the other test sections.

For this analysis, Powers identified test takers who had attended formal coaching programs for any or all of the GRE test sections. For the analytical ability section, the analysis revealed a strong relationship between the effect of coaching and its duration (in terms of hours devoted to instruction). However, applying the same methodology to the verbal and quantitative sections revealed little if any such relationship, contrary to claims being made by commercial coaching firms. Increasing the duration of preparation for the verbal and quantitative GRE measures was not associated with commensurate increases in scores for these two measures.

17.2.2.1 Effects on Relationships of Test Scores with Other Measures

While Messick (1982) provided an insightful *logical* analysis of the ways in which special test preparation may impact validity, there appears to have been little *empirical* research to demonstrate how such practices may affect, for example, the relationship of test scores to other relevant measures. An exception is a study by Powers (1985b), who examined the relationship of GRE analytical ability scores, obtained under ten different randomly assigned test preparation conditions, to indicators of academic performance. Each of the various test preparation conditions was designed, mainly, to help test takers become familiar with each of several novel analytical ability item types. The results suggested that the more time test takers devoted to using the test preparation materials, the stronger the relationship was between academic performance and scores on the GRE analytical ability measure. Specifically, over the ten treatment groups, the correlation between (a) GRE analytical ability score and (b) undergraduate grade point average in the final 2 years of undergraduate study increased according to mean time devoted to preparing for the analytical

measure ($r = .70$, $p < .05$). In addition, correlations of GRE analytical ability scores with GRE verbal and quantitative scores were not significantly related to amounts of test preparation. Thus, both the convergent and (possibly) the discriminant aspects of construct validity of test scores may have been enhanced.

17.3 Summary

ETS has made several contributions to understanding the effects of special test preparation and coaching on (a) test-taking behavior, (b) test performance, and (c) test validity. First, ETS researchers have brought more methodological rigor to the field by demonstrating the feasibility of conducting experimental studies of the effects of test preparation. Rigor has also been increased by introducing more sophisticated methods for controlling self-selection bias in nonexperimental studies.

Moreover, ETS researchers have evaluated the effects of a variety of different types of test preparation: formal commercial coaching, school-offered test preparation programs, and test sponsor-provided test familiarization. With respect to the last type, a significant portion of the ETS-conducted research has focused on making certain that *all* test takers are well prepared, not just those who can afford extensive coaching. Along these lines, researchers have evaluated the effects of test familiarization and other means of test preparation that can be offered, usually remotely for independent study, to *all* test takers. Both secondary and postsecondary student populations have been studied.

Thanks largely to Messick (1980, 1981, 1982), the question of the effectiveness of coaching and test preparation has been reformulated—that is, extended beyond the search for a yes/no answer to the oversimplified question "Does coaching work?" Partly as a result, researchers now seem more inclined to examine the components of test preparation programs in order to ascertain the particular features that are implicated in its effectiveness.

ETS researchers have also stressed that every test is typically composed of a variety of item types and that some of these item types may be more susceptible to coaching and practice than others. In this vein, they have determined some of the features of test item types that seem to render them more or less susceptible. As a consequence, there is now a greater realization that it is insufficient to simply consider the coachability of a test as a whole, but rather it is necessary to consider the characteristics of the various item types that comprise it.

In addition, at least in the view of the scientific community, if not among the general public, a more accurate estimate of the true value of commercial coaching programs now exists. Consumers have information to make more informed choices about whether to seek commercial coaching, for instance. The true effect of coaching on test performance seems neither as negligible as some have claimed nor as large as has been advertised by the purveyors of coaching services.

Most of the studies of coaching and test preparation have focused on the extent to which these practices cause spurious test score improvement. However, although

relatively rare, ETS researchers have also examined, in both a logical and an empirical manner, the effects of test preparation and coaching on the empirical relationships of test scores to other indicators of developed ability.

Finally, ETS research on test preparation has been more than an academic exercise. It has resulted in significant—even dramatic—modifications to several tests that ETS offers. These changes are perhaps the clearest example of the impact of ETS's research on test preparation. However, there have, arguably, been more subtle effects as well. Now, when new assessments are being developed, the potential coachability of proposed new test item types is likely to be a factor in decisions about the final composition of a test. Considerations about test preparation figure into the *design* of tests, well before these tests are ever administered to test takers.

References

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT verbal scores. *American Educational Research Journal, 17*, 239–251. https://doi.org/10.3102/00028312017002239

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*, 373–417. https://doi.org/10.3102/00346543060003373

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429–444). New York: Macmillan.

Bronner, E. (1998, November 24). Study casts doubt on the benefits of S.A.T.-coaching courses. *The New York Times National*, p. A23.

College Entrance Examination Board. (1965). *Effects of coaching on Scholastic Aptitude Test scores*. New York: Author.

Dyer, H. S. (1953). Does coaching help? *College Board Review, 19*, 331–335.

Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10*, 257–272. https://doi.org/10.1111/j.1745-3984.1973.tb00803.x

Faggen, J., & McPeek, M. (1981, April). *Practice effects for four different item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Federal Trade Commission, Boston Regional Office. (1978, September). *Staff memorandum of the Boston Regional office of the Federal Trade Commission: The effects of coaching on standardized admission examinations*. Boston: Author.

Federal Trade Commission, Bureau of Consumer Protection. (1979). *Effects of coaching on standardized admission examinations: Revised statistical analyses of data gathered by Boston Regional office, Federal Trade Commission*. Washington, DC: Author.

French, J. W., & Dear, R. E. (1959). Effect of coaching on an aptitude test. *Educational and Psychological Measurement, 19*, 319–330. https://doi.org/10.1177/001316445901900304

Levy, S. (1979, March). ETS and the coaching cover-up. *New Jersey Monthly*, *3*(4), 50–54, 82–89.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York: Macmillan.

Messick, S. (1980). *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC*. Princeton: Educational Testing Service.

Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias*. San Francisco: Jossey-Bass.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational testing and practice. *Educational Psychologist, 17*, 67–91. https://doi.org/10.1080/00461528209529246

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216. https://doi.org/10.1037/0033-2909.89.2.191

Powers, D. E. (1985a). Effects of coaching on GRE Aptitude Test scores. *Journal of Educational Measurement, 22*, 121–136. https://doi.org/10.1111/j.1745-3984.1985.tb01052.x

Powers, D. E. (1985b). Effects of test preparation on the validity of a graduate admissions test. *Applied Psychological Measurement, 9*, 179–190. https://doi.org/10.1177/014662168500900206

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67–77. https://doi.org/10.1037/0033-2909.100.1.67

Powers, D. E., & Alderman, D. L. (1983). Effects of test familiarization on SAT performance. *Journal of Educational Measurement, 20*, 71–79. https://doi.org/10.1111/j.1745-3984.1983.tb00191.x

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning scores. *Journal of Educational Measurement, 36*, 93–118. https://doi.org/10.1111/j.1745-3984.1999.tb00549.x

Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. *Applied Psychological Measurement, 5*, 141–158. https://doi.org/10.1177/014662168100500201

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76*(2), 266–278. https://doi.org/10.1037/0022-0663.76.2.266

Sesnowitz, M., Bernhardt, K. L., & Knain, D. M. (1982). An analysis of the impact of commercial test preparation courses on SAT scores. *American Educational Research Journal, 19*, 429–441. https://doi.org/10.3102/00028312019003429

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology, 75*, 104–115.

Swinton, S. S., Wild, C. L., & Wallmark, M. M. (1983). *Investigating practice effects on item types in the graduate record examinations aptitude test* (Research Report No. RR-82-56). Princeton: Educational Testing Service.

Wightman, L. E. (1981, April). *GMAT within-test practice effects studies*. Paper presented at the annual meeting of the National Council of Measurement in Education, Los Angeles.

Chapter 18
A Historical Survey of Research Regarding Constructed-Response Formats

Isaac I. Bejar

This chapter chronicles ETS research and development contributions related to the use of constructed-response item formats.[1] The use of constructed responses in testing dates back to imperial China, where tests were used in the selection of civil servants. However, in the United States, the multiple-choice format became dominant during the twentieth century, following its invention and use by the *SAT*® examinations created by the College Board in 1926. When ETS was created in 1947, post-secondary admissions testing was largely based on tests consisting of multiple-choice items. However, from the start, there were two camps at ETS: those who believed that multiple-choice tests were sufficiently adequate for the purpose of assessing "verbal" skills and those who believed that "direct" forms of assessment requiring written responses had a role to play. For constructed-response formats to regain a foothold in American education several hurdles would need to be overcome. Research at ETS was instrumental in overcoming those hurdles.

The first hurdle was that of reliability, specifically the perennial issue of low interrater agreement, which plagued the acceptance of constructed-response formats for most of the twentieth century. The second hurdle was broadening the conception of validity to encompass more than predictive considerations, a process that began with the introduction of construct validity by Cronbach and Meehl (1955). Samuel Messick at ETS played a crucial role in this process by making construct validity relevant to educational tests. An inflexion point in the process of reincorporating constructed-response formats more widely in educational tests was marked

[1] Constructed responses to a prompt or question can range in scope and complexity. Perhaps the most common constructed response is the written essay. However, short written responses to questions are also considered to be constructed, as are spoken answers in response to a prompt, mathematical responses (equations, plotted functions, etc.), computer programs, and graphical responses such as architectural designs.

I.I. Bejar (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: ibejar@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_18

by the publication of *Construction Versus Choice in Cognitive Measurement* (Bennett and Ward 1993), following the indictment of the multiple-choice format by Norm Frederiksen (1984) regarding the format's potentially pernicious influence on education. The chapters in the book made it clear that the choice of format (multiple choice vs. constructed response) includes considerations of validity broadly conceived. Even when there was growing concern about the almost exclusive reliance on the multiple-choice format, there was much more work to be done to facilitate the operational use of constructed-response items since over the preceding decades the profession had come to rely on the multiple-choice format. That work continues to this day at ETS and elsewhere.

Clearly there is more than one way to convey the scope of research and development at ETS to support constructed-response formats. The chapter proceeds largely chronologically in several sections. The first section focuses on the ETS contributions to scoring reliability, roughly through the late 1980s. The next section considers the evolution of validity toward a unitary conception and focuses on the critical contributions by Samuel Messick with implications for the debate around constructed-response formats.

The third section argues that the interest in technology for testing purposes at ETS from early on probably accelerated the eventual incorporation of writing assessment into several ETS admissions tests. That section reviews work related to computer-mediated scoring, task design in several domains, and the formulation of an assessment design framework especially well-suited for constructed-response tasks, evidence-centered design (ECD).

The fourth section describes ETS's involvement in school-based testing, including *Advanced Placement*® (*AP*®), the National Assessment of Educational Progress (NAEP), and the *CBAL*® initiative. A fifth section briefly discusses validity and psychometric research related to constructed-response formats. The chapter closes with some reflections on six decades of research.

18.1 Reliability

The acceptance of the multiple-choice format, after its introduction in the 1926 SAT, together with the growing importance of reliability as a critical attribute of the scores produced by a test, seems to have contributed to the decline of widely used constructed-response forms of assessment in the United States. However, research at ETS was instrumental in helping to return those formats to the assessment of writing in high-stakes contexts. In this section, some of that research is described. Specifically, among the most important ETS contributions are

1. developing holistic scoring
2. advancing the understanding of rater cognition
3. conducting psychometric research in support of constructed responses

Reliability (Haertel 2006) refers to the level of certainty associated with scores from a given test administered to a specific sample and is quantified as a reliability

or generalizability coefficient or as a standard error. However, the first sense of *reliability* that comes to mind in the context of constructed responses is that of *interrater reliability,* or agreement. Unlike responses to multiple-choice items, constructed responses need to be scored by a process (cf. Baldwin et al. 2005) that involves human judgment or, more recently (Williamson et al. 2006), by an automated process that is guided by human judgment. Those human judgments can be more or less fallible and give rise to concerns regarding the replicability of the assigned score by an independent scorer. Clearly, a low level of interrater agreement raises questions about the meaning of scores.

The quantification of interrater disagreement begins with the work of the statistician F. Y. Edgeworth (as cited by Mariano 2002). As Edgeworth noted,

> let a number of equally competent critics independently assign a mark to the (work) … even supposing that the examiners have agreed beforehand as to … the scale of excellence to be adopted … there will occur a certain divergence between the verdicts of competent examiners. (p. 2)

Edgeworth also realized that individual differences among readers could be the source of those errors by noting, for example, that some raters could be more or less severe than others, thus providing the first example of theorizing about rater cognition, a topic to which we will return later in the chapter. Edgeworth (1890) noted,

> Suppose that a candidate obtains 95 at such an examination, it is reasonably certain that he deserves his honours. Still there is an appreciable probability that his real mark, as determined by a jury of competent examiners (marking independently and taking the average of those marks) is just below 80; and that he is pushed up into the honour class by the accident of having a *lenient examiner*. Conversely, his real mark might be just above 80; and yet by accident he might be compelled without honour to take a lower place as low as 63. (emphasis added, p. 470)

The lack of interrater agreement would plague attempts to reincorporate constructed responses into post-secondary admissions testing once multiple-choice items began to supplant them. An approach was needed to solve the interrater reliability problem. A key player in that effort was none other than Carl Brigham (1890–1943), who was the chief architect behind the SAT, which included only multiple-choice items.[2] Brigham was an atypical test developer and psychometrician in that he viewed the administration of a test as an opportunity to experiment and further learn about students' cognition. And experiment he did. He developed an "experimental section" (N. Elliot 2005, p. 75) that would contain item types that were not being used operationally, for example. Importantly, he was keenly interested in the possibility of incorporating more "direct" measures of writing (Valentine 1987, p. 44). However, from the perspective of the College Board, the sponsor of the test, by the 1930s, the SAT was generating significant income, and the Board seemed to have set some limits on the degree of experimentation. According to Hubin (1988),

[2] For a historical account of how Brigham came to lead the development of the SAT, see Hubin (1988).

> the growth of the Scholastic Aptitude Test in the thirties, although quite modest by standards of the next decade, contrasted sharply with a constant decline in applicants for the traditional Board essay examinations. Board members saw the SAT's growth as evidence of its success and increasingly equated such success with the Board's very existence. The Board's perception decreased Brigham's latitude to experiment with the instrument. (p. 241)

Nevertheless, Brigham and his associates continued to experiment with direct measures of writing. As suggested by the following excerpt (Jones and Brown 1935), there appeared to be progress in solving the rater agreement challenge:

> Stalnaker and Stalnaker … present evidence to show that the essay-type test can be scored with rather high reliability if certain rules are followed in formulating questions and in scoring. Brigham … has made an analysis of the procedures used by readers of the English examinations of the College Entrance Examination Board, and believes that the major sources of errors in marking have been identified. A new method of grading is being tried which, he thinks, will lead to greatly increased reliability. (p. 489)

There were others involved in the improvement of the scoring of constructed responses. For example, Anderson and Traxler (1940) argued that[3]

> by carefully formulating the test material and training the readers, it is possible to obtain highly reliable readings of essay examinations. Not only is the reliability high for the total score, but it is also fairly high for most of the eight aspects of English usage that were included in this study. The reliability is higher for some of those aspects that are usually regarded as fairly intangible than for the aspects that one would expect to be objective and tangible. The test makes fair, though by no means perfect, discrimination among the various years of the secondary school in the ability of the pupils to write a composition based on notes supplied to them. The results of the study are not offered as conclusive, but it is believed that, when they are considered along with the results of earlier studies, they suggest that it is highly desirable for schools to experiment with essay-test procedures as means for supplementing the results of objective tests of English usage in a comprehensive program of evaluation in English expression. (p. 530)

Despite these positive results, further resistance to constructed responses was to emerge. Besides reliability concerns, costs and efficiency also were part of the equation. For example, we can infer from the preceding quotation that the scoring being discussed is "analytic" and would require multiple ratings of the same response. At the same time, machine scoring of multiple-choice responses was rapidly becoming a reality[4] (Hubin 1988, p. 296). The potential efficiencies of machine scoring contrasted sharply with the inefficiencies and logistics of human scoring. In fact, the manpower shortages during World War II led the College Board to suspend examinations relying on essays (Hubin 1988, p. 297).

---

[3] In this passage, *reliability* refers to interrater agreement.

[4] Du Bois (1970, p. 119), citing Downey (1965), notes that the scoring machine was invented in 1934 by Reynold B. Johnson, inspired by Ben D. Wood's vision of large-scale testing. According to Du Bois, these scoring machines greatly reduced the cost and "accelerated the trend toward more or less complete reliance on objective tests, especially the multiple-choice item." Of course, testing volume increased over the decades and motivated significant innovations. E. F. Lindquist at the University of Iowa invented the first successful optical scanner in 1962 (U.S. Patent 3,050,248) that was capable of processing larger numbers of answer sheets than the prior electrical mark sense scanner Johnson invented.

However, the end of the war did not help. Almost 10 years on, a study published by ETS (Huddleston 1954) concluded that

> the investigation points to the conclusion that in the light of present knowledge, measurable "ability to write" is no more than verbal ability. It has been impossible to demonstrate by the techniques of this study that essay questions, objective questions, or paragraph-revision exercises contain any factor other than verbal; furthermore, these types of questions measure writing ability less well than does a typical verbal test. The high degree of success of the verbal test is, however, a significant outcome.
>
> The results are discouraging to those who would like to develop reliable and valid essay examinations in English composition—a hope that is now more than half a century old. Improvement in such essay tests has been possible up to a certain point, but professional workers have long since reached what appears to be a stone wall blocking future progress. New basic knowledge of human capacities will have to be unearthed before better tests can be made or more satisfactory criteria developed. To this end the Educational Testing Service has proposed, pending availability of appropriate funds, a comprehensive factor study in which many types of exercises both new and traditional are combined with tests of many established factors in an attempt to discover the fundamental nature of writing ability. The present writer would like to endorse such a study as the only auspicious means of adding to our knowledge in this field. Even then, it appears unlikely that significant progress can be made without further explorations in the area of personality measurement.[5] (pp. 204–205)

In light of the limited conception of both "verbal ability" and "writing ability" at the time, Huddleston's conclusions appear, in retrospect, to be unnecessarily strong and overreaching. The evolving conception of "verbal ability" continues to this day, and it is only recently that even basic skills, like vocabulary knowledge, have become better understood (Nagy and Scott 2000); it was not by any means settled in the early 1950s. Importantly, readily available research at the time was clearly pointing to a more nuanced understanding of writing ability. Specifically, the importance of the role of "fluency" in writing was beginning to emerge (C. W. Taylor 1947) well within the psychometric camp. Today, the assessment of writing is informed by a view of writing as a "complex integrated skill" (Deane et al. 2008; Sparks et al. 2014) with fluency as a key subskill.

By today's standards, the scope of the concept of *reliability* was not fully developed in the 1930s and 1940s in the sense of understanding the *components* of unreliability. The conception of reliability emerged from Spearman's work (see Stanley 1971, pp. 370–372) and was focused on *test-score* reliability. If the assignment of a score from each component (item) is error free, because it is scored objectively, then the scoring does not contribute error to the total score, and in that case score reliability is a function of the number of items and their intercorrelations. In the case of constructed responses, the scoring is not error free since the scorer renders a judgment, which is a fallible process.[6] Moreover, because items that require constructed responses require more time, typically, fewer of them can be administered which, other things

[5] The mysterious reference to "personality measurement" appears to be reference to the thinking that personality measurement would be the next frontier in admissions testing. In fact, ETS, specifically Henry Chauncey, was interested in personality measurement (see Lemann 1999, p. 91).

[6] Fallibility is relative; even the scoring of multiple-choice items is not 100% error free, at least not without many preventive measures to make it so. For a discussion, see Baker (1971).

being equal, reduces score reliability. The estimation of error components associated with ratings would develop later (Ebel 1951; Finlayson 1951), as would the interplay among those components (Coffman 1971), culminating in the formulation of generalizability theory (Cronbach et al. 1972).[7] Coffman (1971), citing multiple sources, summarized the state of knowledge on interreader agreement as follows:

> The accumulated evidence leads, however, to three inescapable conclusions: a) different raters tend to assign different grades to the same paper; b) a single rater tends to assign different grades to the same paper on different occasions; and c) the differences tend to increase as the essay question permits greater freedom of response. (p. 277)

Clearly, this was a state of affairs not much different than what Edgeworth had observed 80 years earlier.

18.1.1 The Emergence of a Solution

The Huddleston (1954) perspective could have prevailed at ETS and delayed the wider use of constructed responses, specifically in writing.[8] Instead, from its inception ETS research paved the way for a solution to reducing interrater disagreement. First, a groundbreaking investigation at ETS (funded by the Carnegie Corporation) established that raters operated with different implied scoring criteria (Diederich et al. 1961). The investigation was motivated by the study to which Huddleston refers in the preceding quotation. That latter study did not yield satisfactory results, and a different approach was suggested: "It was agreed that further progress in grading essays must wait upon a factor analysis of judgments of a diverse group of competent readers in an unstructured situation, where each could grade as he liked" (Diederich et al. 1961, p. 3). The motivating hypothesis was that different readers belong to different "schools of thought" that would presumably value qualities of writing differently. The methodology that made it possible to identify types of readers was first suggested by Torgerson and Green (1952) at ETS. To identify the schools of thought, 53 "distinguished readers" were asked to rate and annotate 300 papers without being given standards or criteria for rating. The factors identified from the interrater correlations consisted of groupings of raters (e.g., raters that loaded highly on a specific factor). What school of thought was represented by a given factor would not be immediately obvious without knowing the specifics of the reasoning underlying a rater's judgment. The reasoning of the readers was captured by means of the annotations each judge had been asked to make, which then had to be coded and classified.[9] The results showed that agreement among readers was

[7] Brennan (2001, p. 3) credits Burt in 1936 and Lindquist in 1953 with anticipating the essence of univariate generalizability theory.

[8] See Diederich (1957) for a candid description of the state of affairs with respect to using essays in admissions testing.

[9] A very laborious process carried out by Sydell Carlton, an employee at ETS until 2017. She recalls (personal communication, July 19, 2010) that no one could initially interpret the factors.

poor and that the nature of the schools of thought was that they valued different aspects of writing. However, the two most sharply defined groups were those that valued "ideas" or that valued "mechanics."

The Diederich et al. (1961) study showed that judges, when left to their own analytical devices, will resort to particular, if not idiosyncratic, evaluative schemes and that such particularities could well explain the perennial lack of adequate interrater agreement. Important as that finding was, it still did not formulate a solution to the problem of lack of interrater agreement. That solution took a few more years, also leading to a milestone in testing by means of constructed responses. The study was carried out at ETS and led by Fred I. Godshalk. The study was ambitious and included five 20-minute essays, six objective tests, and two interlinear exercises, administered to 646 12th graders over a period of several weeks. Importantly, the scoring of the essays was *holistic*. They defined the scoring procedure of the essays as follows (Godshalk et al. 1966):

> The readers were asked to make global or holistic, not analytical, judgments of each paper, *reading rapidly for a total impression*. There were only three ratings: a score of "3" for a superior paper, "2" for an average paper, and "1" for an inferior paper. The readers were told to judge each paper on its merits without regard to other papers on the same topic; that is, they were not to be concerned with any ideas of a normal distribution of the three scores. They were advised that scores of "3" were possible and that the "safe" procedure of awarding almost all "2s" was to be avoided. Standards for the ratings were established in two ways: by furnishing each reader with copies of the sample essays for inspection and discussion, and by explaining the conditions of administration and the nature of the testing population; and by having all readers score reproduced sets of carefully selected sample answers to all five questions and to report the results. The scores were then tabulated and announced. No effort was made to identify any reader whose standards were out of line, because that fact would be known to him and would be assumed to have a corrective effect. The procedure was repeated several times during the first two days of scoring to assist readers in maintaining standards. (p. 10, emphasis added)

Perhaps the critical aspect of the directions was to "to make global or holistic, not analytical, judgments" and the use of what is known today (Baldwin et al. 2005) as benchmark or range finding papers to illustrate the criteria. The authors describe the procedure in the preceding quotation and do not provide a theoretical rationale. They were, of course, aware of the earlier Diederich study, and it could have influenced the conception of the holistic scoring instructions. That is, stressing that the scoring was to be holistic and not analytical could have been seen as way to prevent the schools of thought from entering the scoring process and to make the scoring process that much faster.[10]

After she classified a few of the annotations, she formulated a coding scheme that could be used to systematically annotate the rest of essays. The actual coding of more than 10,000 papers was hired out. By examining the annotation of readers that loaded highly on one factor or another, it became possible to interpret the factors as schools of thought.

[10] Although Godshalk et al. (1966) are associated with making *holistic* scoring a widely accepted approach, the term "wholistic" was used first at ETS by Ann F. Coward with the same meaning. In a project published as a brief internal report (Coward 1950) that was subsequently published (Coward 1952), she compared "wholistic," which corresponded with what later was called *holistic*,

Outside of ETS, the development of holistic scoring was well received by teachers of English (White 1984) and characterized as "undoubtedly one of the biggest breakthroughs in writing assessment" (Huot 1990, p. 201). Interestingly, other concurrent work in psychology, although relevant in retrospect, was not considered at the time as related to scoring of essays. For example, N. Elliot (2005) postulated the relevance of Gestalt psychology to a possible adoption of holistic scoring, although there is no such evidence in the Godshalk et al. (1966) report. Another line of research that was relevant was models of judgment, such as the lens model proposed by Egon Brunswik (Brunswik 1952; Hammond et al. 1964; Tucker 1964).[11] The lens model, although intended as a perceptual model, has been used primarily in decision-making (Hammond and Stewart 2001). According to the model, the perceiver or decision maker decomposes an object into its attributes and weighs those attributes in arriving at a judgment. The model is clearly applicable in modeling raters (Bejar et al. 2006). Similarly, a theory of personality of the same period, George Kelly's personal construct theory, included a method for eliciting "personal constructs" by means of analysis of sets of important others.[12] The method, called the repertory grid technique, was later found useful for modeling idiographic or reader-specific rating behavior (Bejar et al. 2006; Suto and Nadas 2009).

One additional area of relevant research was the work on clinical judgment. Meehl's (1954) influential monograph concluded that actuarial methods were superior to clinical judgment in predicting clinical outcomes. One reason given for the superiority of actuarial methods, often implemented as a regression equation or even the sum of unweighted variables (Dawes and Corrigan 1974), is that the actuarial method is provided with variables from which to arrive at a judgment. By contrast, the clinician first needs to figure out the variables that are involved, the rubric, so to speak, and determine the value of the variables to arrive at a judgment. As Meehl stressed, the clinician has limited mental resources to carry out the task. Under such conditions, it is not unreasonable for the clinician to perform inconsistently relative to actuarial methods. The overall and quick impression called for by the holistic instructions could have the effect of reducing the cognitive load demanded by a very detailed analysis. Because of this load, such an analysis is likely to play upon the differences that might exist among readers with respect to background and capacity to carry out the task.

There was such relief once the holistic method had been found to help to improve interrater agreement that no one seems to have noted that the idea of holistic scoring is quite counterintuitive. How can a quick impression substitute for a deliberate and

---

and "atomistic" approaches to scoring. No great differences between the two methods were reported, nor was any rationale proposed for the "wholistic" method. There was also experimentation in the UK on impressionistic scoring in the early 1960s (N. Elliot, personal communication, May 15, 2015)

[11] Ledyard Tucker, the eminent ETS psychometrician, had been a reviewer of the Hammond et al. (1964) paper. His review so influenced the Hammond et al. paper that Hammond suggested to the *Psychological Review* editors that Tucker's formulation of the lens model appear as an independent paper (Hammond, personal communication, March 29, 2010).

[12] George Kelly's work was well known at ETS (Messick and Kogan 1966).

extensive analysis of a constructed response by a subject matter *expert*? Research on decision making suggests, in fact, that experts operate in a holistic sort of fashion and that it is a sign of their expertise to do so. Becoming an expert in any domain involves developing "fast and frugal heuristics" (Gigerenzer and Goldstein 1996) that can be applied to arrive at accurate judgments quickly.

Eventually, questions would be raised about holistic scoring, however. As Cumming et al. (2002) noted,

> holistic rating scales can conflate many of the complex traits and variables that human judges of students' written compositions perceive (such as fine points of discourse coherence, grammar, lexical usage, or presentation of ideas) into a few simple scale points, rendering the meaning or significance of the judges' assessments in a form that many feel is either superficial or difficult to interpret. (p. 68)

That is, there is a price for the increased interreader agreement made possible by holistic scoring, namely, that we cannot necessarily document the mental process that scorers are using to arrive at a score. In the absence of that documentation, strictly speaking, we cannot be sure by what means scores are being assigned and whether those means are appropriate until evidence is presented.

Concerns such as these have given rise to research on rater cognition (Bejar 2012). The Diederich et al. (1961) study at ETS started the research tradition by attempting to understand the basis of lack of agreement among scorers (see also Myers et al. 1966). The range of the literature, a portion of it carried out at ETS, is vast and aims, in general, to unpack what goes on in the minds of the raters as they score (Bejar et al. 2006; Crisp 2010; Elbow and Yancey 1994; Huot and Neal 2006; Lumley 2002; Norton 1990; Pula and Huot 1993; Vaughan 1991), the effect of a rater's background (Myford and Mislevy 1995; Shohamy et al. 1992), rater strategies (Wong and Kwong 2007), and methods to elicit raters' personal criteria (Bejar et al. 2006; Heller et al. 1998). Descriptions of the qualifications of raters have also been proposed (Powers et al. 1998; Suto et al. 2009). In addition, the nature of scoring expertise has been studied (Wolfe 1997; Wolfe et al. 1998). Methods to capture and monitor rater effects during scoring as a function of rater characteristics are similarly relevant (Myford et al. 1995; Myford and Mislevy 1995; Patz et al. 2002). Experimental approaches to modeling rater cognition have also emerged (Freedman and Calfee 1983), where the interest is on systematic study of different factors that could affect the scoring process. The effectiveness of different approaches to the training of readers (Wolfe et al. 2010) and the qualifying of raters (Powers et al. 1998) has also been studied. In short, the Diederich et al. study was the first in a long line of research concerned with better understanding and improving the processes in which raters engage.

A second concern regarding holistic scoring is the nature of the inferences that can be drawn from scores. Current rubrics described as holistic, such as those used for scoring the *GRE*® analytical writing assessment, are very detailed, unlike the early rubrics. That is, holistic scoring has evolved from its inception, although quietly. Early holistic scoring had as a goal the *ranking* of students' responses.

Holistic scoring emerged in the context of admissions testing, which means in a norm-referenced context. In that context, the ranking or comparative interpretation of candidates is the goal. Points along the scale of such a test do not immediately have implications for what a test taker knows and can do, that is, attaching an interpretation to a score or score range. The idea of criterion-referenced (Glaser 1963) measurement emerged in the 1960s and was quickly adopted as an alternative conception to norm-referenced testing, especially in the context of school-based testing. Today it is common (Linn and Gronlund 2000) to talk about standards-based assessments to mean assessments that have been developed following a framework that describes the content to be assessed such that scores on the test can be interpreted with respect to what students know and can do. Such interpretations can be assigned to a single score or, more commonly, a range of scores by means of a process called standard setting (Cizek and Bunch 2007; Hambleton and Pitoniak 2006), where panels of experts examine the items or performance on the items to determine what students in those score regions know and can do.

NAEP had from its inception a standards-based orientation. The initial implementation of NAEP in the 1960s, led by Ralph Tyler, did not report scores, but rather performance on specific items, and did not include constructed responses. When writing was first introduced in the late 1960s, the scoring methodology was holistic (Mullis 1980, p. 2). However, the methodology was not found adequate for NAEP purposes and instead the method of primary traits was developed for the second NAEP writing assessment in 1974 (Cooper 1977, p. 11; Lloyd-Jones 1977). The inapplicability of holistic scoring to NAEP measurement purposes is given by Mullis (1980):

> NAEP needed to report performance levels for particular writing skills, and the rank ordering did not readily provide this information. Also, NAEP for its own charge of measuring change over time, as well as for users interested in comparisons with national results, needed a scoring system that could be replicated, and this is difficult to do with holistic scoring. (p. 3)

The criterion-referenced rationale that Mullis advocated was very much aligned with the standards-based orientation of NAEP. According to Bourque (2009), "by the mid-1980s, states began to realize that better reporting mechanisms were needed to measure student progress" (p. 3). A policy group was established, the National Assessment Governing Board (NAGB), to direct NAEP, and shortly thereafter the "Board agreed to adopt three achievement levels (Basic, Proficient, and Advanced) for each grade and subject area assessed by NAEP" (Bourque 2009, p. 3).

With respect to writing, Mullis (1984) noted,

> For certain purposes, the most efficient and beneficial scoring system may be an adaptation or modification of an existing system. For example, the focused holistic system used by the Texas Assessment Program … can be thought of as a combination of the impressionistic holistic and primary trait scoring systems. (p. 18)

To this day, the method used by NAEP to score writing samples is a modified holistic method called focused holistic (H. Persky, personal communication, January 25,

2011; see also, Persky 2012) that seems to have first originated in Texas around 1980 (Sachse 1984).

Holistic scoring also evolved within admissions testing for different reasons, albeit in the same direction. N. Elliot (2005, p. 228) gives Paul Ramsey at ETS credit for instituting a "modified holistic" method to mean that the scoring was accompanied by detailed scoring guides. In 1992 the College Board's English Composition Test (which would become SAT Writing) began using scoring guides as well. The rationale, however, was different, namely, comparability:[13]

> We need a scoring guide for the SAT Writing test because, unlike the ECT [English Composition Test] which gives an essay once a year, the SAT will be given 5 times a year and scoring of each administration must be comparable to scoring of other administrations. Other tests, like TOEFL, which give an essay several times a year use a scoring guide like this. (Memorandum from Marylyn Sudlow to Ken Hartman, August 6, 1992)

Clearly the approach to scoring of constructed responses had implications for score meaning and score comparability. However, the psychometric support for constructed responses was limited, at least compared with the support available for multiple-choice tests. Psychometric research at ETS since the 1950s was initially oriented to dichotomously scored items; a historical account can be found in Carlson and von Davier (Chap. 5, this volume). Fred Lord's work (Lord 1952) was critical for developing a broadly applicable psychometric framework, item response theory (IRT), that would eventually include ordered polytomously scored items (Samejima 1969),[14] a needed development to accommodate constructed responses. Indeed, IRT provided the psychometric backbone for developing the second generation of NAEP (Messick et al. 1983), including the incorporation of polytomously scored constructed-response items at a time when to do so in large-scale testing was rare. (For a detailed discussion of the ETS contributions to psychometric theory and software in support of constructed-response formats, see Carlson and von Davier, Chap. 5, this volume.)

The sense of error of measurement within IRT, as represented by the idea of an information function (Birnbaum 1968), was conditional and sample independent (in a certain sense), an improvement over the conception of error in classical test theory, which was global and sample specific. IRT introduced explicitly the idea that the error or measurement was not constant at all ability levels, although it did not allow for the identification of sources of error. Concurrent developments outside the IRT sphere made it possible to begin teasing out the contribution of the scoring process to score reliability (Ebel 1951; Finlayson 1951; Lindquist 1953), culminating in generalizability theory (Cronbach et al. 1972). Such analyses were useful for

[13] Interestingly, the rationale underlying Sudlow's memorandum is the same as the rationale for instituting the methodology of equating in the SAT itself in the 1940s (College Entrance Examination Board 1942, p. 34), namely, that the SAT would be administered more than once per year and the two within-year testing populations could not be assumed to be equivalent as a year-to-year population might be.

[14] Fumiko Samejima was at ETS during the 1960s, invited by Fred Lord. A full account can be found in Wainer and Robinson (2007).

characterizing what portion of the error variability was due to different sources, among them lack of reader agreement. Bock et al. (2002), however, proposed a solution to incorporate that framework into IRT whereby the conditional standard error of measurement derived from IRT could be partitioned to identify the portion due to the rating process. (Briggs and Wilson 2007, provide for a more elaborate integration of IRT and generalizability theory.)

As had been recognized by Edgeworth (1890), readers can differ in the stringency of the scores they assign and such disagreements contribute to the error of measurement. Henry Braun[15] appears to have been the first one at ETS to introduce the idea of rater calibration, described earlier by Paul (1981), as an approach to compensate for systematic disagreements among raters. The logic of the approach was described as follows (Braun 1988): "This new approach involves appropriately adjusting scores in order to remove the noise contributed by systematic sources of variation; for example, a reader consistently assigning higher grades than the typical reader. Such adjustments are akin to an equating process" (p. 2).

The operational implementation of the idea would prove challenging, however. To implement the idea economically, specialized data collection designs were necessary and needed to be embedded in the operational scoring process over several days. The effects estimated from such an analysis are then used to adjust the raw scores. Along the same lines, Longford (1994) also studied the possibility of adjusting scores by taking into account rater severity and consistency.

An alternative to adjusting scores retrospectively is to identify those readers who appear to be unusually severe or lenient so that they can receive additional training. Bejar (1985) experimented with approaches to identify "biased" readers by means of multivariate methods in the Test of Spoken English. Myford et al. (1995) approached the problem of rater severity by applying FACETS (Linacre 2010), an extension of the IRT Rasch model that includes rater parameters, as well as the parameters for test takers and items.

18.1.2 Conclusion

When ETS was formed, the pragmatics of increasingly large scale testing together with psychometric considerations set a barrier to the use of constructed-response formats, which was viewed as unreliability due to inadequate interrater agreement. Carl Brigham, chief developer of the SAT, was also a strong proponent of more direct measures, but a solution to the scoring problem eluded him. After Brigham's death, there appeared to be no strong proponent of the format, at least not within the College Board, nor in the initial years of ETS. Without Brigham to push the point, and the strong undercurrent against constructed responses illustrated by Huddleston's (1954) perspective that writing skills do not merit their own construct, the prospects

[15] Henry Braun was vice president for research management from 1990 to 1999.

for constructed-response testing seemed dire. However, the ETS staff also included writing scholars such as Paul Diederich and Fred Godshalk, and because of them, and others, ultimately there was significant progress in solving the interrater agreement challenge with the emergence of holistic scoring. That method, which was also widely accepted outside of ETS paved the way for an increase in the use of essays. However, as we will see in the next sections, much more was needed for constructed-response formats to become viable.

18.2 Validity

Making progress on the scoring of constructed responses was critical but far from sufficient to motivate a wider reliance on constructed-response formats. Such formats necessarily require longer response times, which means fewer items can be administered in a given time, threatening *score* reliability. The conception of validity prevailing in the mid-twentieth century emphasized predictive validity, which presented a challenge for the adoption of constructed-response formats since their characteristic lower score reliability would attenuate predictive validity. The evolution of validity theory would be highly relevant to decisions regarding the use of response format, as we will see shortly. Research at ETS played a key role and was led by Samuel Messick, who not only would argue, along with others, for a unitary—as opposed to a so-called Trinitarian—conception of validity (Guion 1980) (consisting of content, criterion and construct "validities") but also, as important, for the relevance of such a unitary conception of validity to *educational* measurement. First, it is informative to review briefly the historical background.

The notion that eventually came to be known as *content validity*, and was seen as especially relevant to educational testing, probably has its roots in the idea of the sampling of items as a warrant for score interpretation. That notion was proposed early on by Robert C. Tryon as a reaction to the factor analytic conception of individual differences that prevailed at the time. Tryon (1935) argued,

> The significant fact to observe about mental measurement is that, having marked out by definition some domain for testing, the psychologist chooses as *a method of measurement* one which indicates that he knows *before giving the test to any subjects* a great deal about the nature of the factors which cause individual differences in the domain. The method is that of *sampling* behavior, and it definitely presupposes that for any defined domain there exists a *universe* of causes, or factors, or components determining individual differences. Each test-item attempts to 'tap' one or more of these components. (p. 433, emphasis in the original)

Tryon was on track with respect to assessment design by suggesting that the assessment developer should know much about what is to be tested "*before giving the test to any subject*," therefore implying the need to explicate what is to be measured in some detail as a first step in the design of an assessment (a principle fully fleshed out in ECD, Mislevy et al. 2003, much later). However, his rejection of the prevailing factor analytic perspective advocated by the prominent psychologists of

the day (Spearman 1923; Thurstone 1926) was probably responsible for the lack of acceptance of his perspective.[16] Among the problems raised about the sampling perspective as a warrant to score interpretation was that, in principle, it seemed to require the preexistence of a universe of items, so that random samples could be taken from it. Such an idea presupposes some means of defining the universe of items. The resistance to the idea was most vocally expressed by Jane Loevinger (1965), who could not envision how to explicate such universes. Nevertheless, the relevance of sampling in validation was affirmed by Cronbach (1980) and Kane (1982), although not as a sufficient consideration, even though the link back to Tryon was lost along the way.

What appears to have been missed in Tryon's argument is that he intended the universe of items to be isomorphic with a "*universe* of factors, causes, or components determining individual differences" (p. 433), which would imply a crossing of content and process in the creation of a universe of items. Such an idea foreshadows notions of validity that would be proposed many decades later, specifically notions related to construct representation (Embretson 1983). Instead, in time, the sampling perspective became synonymous with content validity (Cronbach 1971): "Whether the operations that finally constitute the test correspond to the *specified universe* is the question of content validity" (p. 452, emphasis added). The idea of a universe was taken seriously by Cronbach (although using for illustration an example from social psychology, which, interestingly, implies a constructed-response test design):

> For observation of sociability, the universe specification presumably will define a category of "social acts" to be tallied and a list of situations in which observations are to be made. Each observation ought to have validity as a sample from this universe. (p. 452)

While sampling considerations evolved into content validity, and were thought to be especially applicable to educational (achievement) testing (Kane 2006), the predictive or criterion notion of "validity" dominated from 1920 to 1950 (Kane 2006) and served to warrant the use of tests for selection purposes, which in an educational context meant admissions testing. The research at ETS described earlier on writing assessment took place in that context. The predictive view presented a major hurdle to the use of constructed-response formats because, in a predictive context, it is natural to evaluate any modifications to the test, such as adding constructed-response formats, with respect to *increases* in prediction (Breland 1983):

> Because of the expense of direct assessments of writing skill, a central issue over the years has been whether or not an essay adds significantly to the measurement accuracy provided by other available measures-the high school record, objective test scores, or other information. (p. 14)

Breland provided a meta-analysis of writing assessment research showing the incremental prediction of writing samples over measures consisting only of

[16] The caution about factor analysis was expressed many decades later by Sam Messick (1972): "these concerns [about factor analysis] could lead to a marked skepticism about the construct validity of empirically derived factors as fundamental dimensions of behavioral processes" (p. 358).

multiple-choice items. Although he presented a fairly compelling body of evidence, a cost-conscious critic could have argued that the increases in prediction could just as easily have been obtained more economically by lengthening the multiple-choice component.

The third conception of validity is construct validity, dating back to the mid-twentieth-century seminal paper introducing the term (Cronbach and Meehl 1955). In that paper, validation is seen as a process that occurs after the assessment has been completed, although the process is driven by theoretical expectations. However, Cronbach and Meehl did not suggest that those expectations should be used in developing the test itself. Instead, such theoretical expectations were to be used to locate the new test within a nomological network of relationships among theoretically relevant variables and scores. At the time Cronbach and Meehl were writing, developing a test was a matter of writing items as best one could and then pretesting them. The items that did not survive were discarded. In effect, the surviving items were the de facto definition of the construct, although whether it was the intended construct could not be assumed until a conclusion could be reached through validation. In the wrong hands, such an ad hoc process could converge on the wrong test.[17] Loevinger (1957) argued that "the dangers of pure empiricism in determining the content of a test should not be underestimated" (p. 657) and concluded that

> there appears to be no convincing reason for ignoring content nor for considering content alone in determining the validity of a test or individual items. The problem is to find a coherent set of operations permitting utilization of content together with empirical considerations. (p. 658)

Clearly Loevinger considered content important, but the "coherent set of operations" she referred to was missing at the time, although it would appear soon as part of the cognitive science revolution that was beginning to emerge in the 1950s.[18]

Toward the end of that decade, another important article was published that would have important repercussions for the history of research on constructed-response formats. D. T. Campbell and Fiske (1959) made an important distinction: "For the justification of novel trait measures, for the validation of test interpretation, or for the establishment of construct validity, *discriminant* validation as well as convergent validation is required" (p. 81, emphasis in the original).

The paper is significant for contrasting the evidentiary basis for *and* against a psychometric claim.[19] In addition, the paper formalizes the notion of *method*

[17] Of course, in the right hands, the approach could also converge on a very effective instrument. Two of the most highly regarded assessments developed during the twentieth century, the Minnesota Multiphasic Personality Inventory (MMPI) and the Strong Vocational Interest Blank, later to become the Strong–Campbell, were developed in this fashion.

[18] Miller (2003), a major leader of the revolution, provides a historical account of cognitive science.

[19] Toulmin's (1958) model of argument was published at around the same time. Toulmin stressed the role of counterarguments and rebuttals of claims or conclusions. Toulmin's model figured prominently in the subsequent evolution of validation (Kane 2006) and in assessment design (Mislevy et al. 2003). Karl Popper's (1959/1992) book, *The Logic of Scientific Discovery,* also appeared in 1959. Popper stressed the importance of falsifying theories, a concept that can be

variance, which would surface later in research about constructed-response formats, especially evaluating the measurement equivalence of multiple-choice and constructed-response formats.

As can be seen, the 1950s was a contentious and productive decade in the conceptual development of testing. Importantly, the foregoing discussion about the nature of validity did *not* take place at ETS. Nevertheless, it is highly relevant to the chapter: These developments in validity theory may have even been seen as tangential to admissions tests,[20] which represented the vast majority of ETS operations at the time. In that context, the normative interpretations of scores together with predictive validity were the accepted practice.

As mentioned earlier, in 1963 a most influential paper was published by Glaser, proposing an alternative approach to score interpretation and assessment design in an educational setting, namely, by *reference* to the level of proficiency within a very well-defined content domain. Glaser's intent was to provide an alternative to normative interpretations since norms were less relevant in the context of individualized instruction.[21] Whereas norms provide the location of a given score in a distribution of scores, a criterion-referenced interpretation was intended to be more descriptive of the test taker's skills than a normative interpretation. The criterion-referenced approach became aligned early on with the idea of mastery testing (Hambleton and Novick 1973), whereby the objective of measurement was to determine whether a student had met the knowledge requirements associated with a learning objective.

Criterion-referenced tests were thought to yield more actionable results in an educational context not by considering a score as a deviation from the mean of a distribution, the normative interpretation, but by locating the score within an interval whereby all scores in that interval would have a similar interpretation. In the simplest form, this meant determining a cut score that would define the range of pass scores and the range for fail scores, with pass implying mastery. To define those intervals, cut scores along the score scale needed to be decided on first. However, as noted by Zieky (1995), the methodology for setting such cut scores had not yet emerged. In retrospect, it is clear that if the deviation from a mean was not adequate for score interpretation, locating a score within an interval would not necessarily help either; much more was needed. In fact, reflecting on his 1963 paper, Glaser (1994) noted that "systematic techniques needed to be developed to more adequately identify and describe the components of performance, and to determine the relative weighting of these components with respect to a given task" (p. 9).

---

applied to challenge assertions about the validity of scores. Messick (1989) discussed Toulmin and Popper at length.

[20] An examination of the titles of research published in ETS's first decades clearly emphasizes predictive validity. However, consequential implications of testing or test bias appeared in the mid-1960s with the work of T. Anne Cleary (1966) and even earlier (Turnbull 1949). Also, Gulliksen's (1950) idea of intrinsic validity, cited by Cronbach and Meehl (1955), is a rare exception on early validity theorizing at ETS, to be followed some years later by the seminal theoretical work of Messick. See Kane and Bridgeman Chap. 16, (this volume) for a comprehensive description of Messick's work and a historical review of validity theory at ETS more generally.

[21] Glaser credits Ebel (1962), who was vice president at ETS at the time, with a similar idea.

The "components of performance" that Glaser thought needed to be developed echoed both Tryon's earlier "components determining individual differences" and the "coherent set of operations permitting utilization of content" that Loevinger called for. That is, there had been an implied consensus all along as to a key ingredient for test meaning, namely, identifying the underlying sources of variability in test performance, which meant a deeper understanding of the response process itself.[22]

18.2.1 Validity Theory at ETS

With the benefit of hindsight, it seems that by 1970, the conception of validity remained divided, consisting of different "validities," which had significant implications for the use of constructed-response formats in education. Three developments were needed to further that use:

- With criterion (and especially, predictive) validity as a primary conception of validity, economics would delay wider use of constructed-response formats. Replacing the Trinitarian view with a unitary view was needed to avoid associating the different "validities" with specific testing contexts.
- Even under a unitary view, the costs of constructed-response formats would remain an obstacle. An expansion of the unitary conception was necessary to *explicitly* give the evidential *and* consequential aspects of validity *equal* footing. By doing so, the calculus for the deployment of constructed-response formats would balance monetary cost with the (possibly intangible) benefits of using the format.
- As alluded to earlier, by and large, the broader discussion of validity theory was not directed at educational achievement testing. Thus, the third needed development was to make the evolution of validity theory applicable to educational testing.

These developments were related and *enormous*. Unlike the earlier evolution of validity, which had taken place outside of ETS, Sam Messick dedicated two decades to explicating the unitary view, bringing its evidential and consequential aspects more into line with one another, and making the view relevant, if not central, to educational testing. These advances, arguably, were essential to wider use of constructed-response formats in education.

Calls for a unitary view in the form of construct validity began early on. Messick (1989) quoted Loevinger that, "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 17). Messick elaborated that idea, stating that, "almost any kind

[22] Of course, generalizability theory had been under development (Rajaratnam et al. 1965) during the 1960s, and it was concerned with components of *observed* score variability. It distinguishes between components of variability that attenuate the interpretation of a score, that is, error variability, and true score variability, summarizing the results into a generalizability coefficient. It does not address the understanding of the response process.

of information about a test can contribute to an understanding of construct validity, but the contribution becomes stronger if the degree of fit of the information with the theoretical rationale underlying score interpretation is explicitly evaluated" (p. 17). That is, Messick stressed the need for a theoretical rationale to integrate the different sources of validity evidence.

Importantly, Messick's (1980, 1989) unitary view explicitly extended to the consequences of test use, with implications for the use of constructed-response formats. Although the message was not well received in some quarters (Kane 2006, p. 54), it was in others. For example, Linn et al. (1991) argued,

> If performance-based assessments are going to have a chance of realizing the potential that the major proponents in the movement hope for, it will be essential that the consequential basis of validity be given much greater prominence among the criteria that are used for judging assessments. (p. 17)

By the 1990s, there had been wider acceptance that consequential evidence was relevant to validity. But that acceptance was one of the later battles that needed to be fought. The relevance of the unitary view to educational testing needed to be established first. In 1975, Messick wondered, "Why does educational measurement, by and large, highlight comparative interpretations, whether with respect to norms or to standards,[23] and at the same time play down construct interpretations?" (p. 957).

This question was raised in reaction to the predominance that criterion-referenced testing had acquired by the 1970s, Among the possible answers Messick (1975) proposed for the absence of construct interpretations was the "legacy of behaviorism and operationism that views desired behaviors as ends in themselves with little concerns for the *processes* that produce them" (p. 959, emphasis added). That speculation was later corroborated by Lorie Shepard (1991) who found that, for the most part, state testing directors had a behaviorist conception of student learning.

The positive attitude toward behaviorism among state testing directors is informative because the so-called cognitive revolution had been under way for several decades. Although its relevance was recognized early on, its impact on testing practice was meager. Susan Embretson, who was not associated with ETS, recognized those implications (Whitely and Dawis 1974).[24] In an important paper, Embretson integrated ideas from cognitive science into testing and psychometric theory by building on Loevinger's argument and layering a cognitive perspective on it. Embretson (1983) proposed the term *construct representation* to describe the extent to which performance on a test is a function of mental processes hypothesized to underlie test performance. An approach to documenting construct representation is modeling the difficulty of items as a function of variables representing the response process and knowledge hypothesized to underlie performance.[25] Modeling of item

[23] In the quotation, by "standards," he meant criterion referencing.

[24] Susan Embretson published as Susan Whitely earlier in her career.

[25] For example, the classic item type based on verbal analogies was thoroughly reanalyzed from a cognitive perspective (Bejar et al. 1991; Pellegrino and Glaser 1980; Sternberg 1977; Whitely and Dawis 1974) with the goal of understanding the variability in the difficulty of the items as a function of the process and knowledge assumed to be involved in analogical reasoning.

difficulty is well suited to multiple-choice items, but less so for items requiring a constructed response since there would typically be fewer of them in any given test. Nevertheless, the concept is equally applicable, as Messick (1994) noted with specific reference to performance assessment: "Evidence should be sought that the presumed sources of task complexity are indeed reflected in task performance and that the complex skill is captured in the test scores with minimal construct underrepresentation" (p. 20).

Embretson's construct representation fit well with Messick's calls for a fuller understanding of the response process as a source of validity evidence. But Messick (1990) also understood that format had the potential to introduce irrelevancies:

> Inferences must be tempered by recognizing that the test not only samples the task universe but casts the sampled tasks in a test format, thereby raising the specter of context effects or irrelevant method [i.e., format] variance possibly distorting test performance vis-a-vis domain performance. (p. 9)

Independently of the evolution of validity theory that was taking place, the calls for direct and authentic forms of assessment never stopped, as evidenced by the work on portfolio assessments at ETS (Camp 1993; Gentile 1992; Myford and Mislevy 1995) and elsewhere. Following the period of "minimum competency testing" in the 1980s there were calls for testing higher order forms of educational achievement (Koretz and Hamilton 2006), including the use of so-called authentic assessments (Wiggins 1989). The deployment of highly complex forms of assessment in the early 1990s was intended to maximize the positive educational consequences of constructed-response formats and avoid the negative consequences of the multiple-choice format, such as teaching to the narrow segment of the curriculum that a multiple-choice test would represent. However, despite the appeal of constructed-response formats, such forms of assessment still needed to be evaluated from a validity perspective encompassing both evidential and consequential considerations. As Messick (1994) noted,

> some aspects of all testing, even performance testing, may have adverse as well as beneficial educational consequences. And if both positive and negative aspects, whether intended or unintended, are not meaningfully addressed in the validation process, then the concept of validity loses its force as a social value. (p. 22)

Indeed, following the large-scale deployment of performance assessments in K–12 in the 1990s (Koretz and Hamilton 2006), it became obvious that overcoming the design challenges would take time. Although the assessments appeared to have positive effects on classroom practice, the assessments did not meet technical standards, especially with respect to score reliability. As a result, the pendulum swung back to the multiple-choice format (Koretz and Hamilton 2006, p. 535).

Not surprisingly, after the long absence of constructed-response formats from educational testing, the know-how for using such formats was not fully developed. Reintroducing such formats would require additional knowledge and a technological infrastructure that would make the format affordable.

18.2.2 Conclusion

Arguably, the predictive conception of validity prevalent through most of the twentieth century favored the multiple-choice format. The evolution of validity into a more unitary concept was not seen initially as relevant to educational measurement. Samuel Messick thought otherwise and devoted two decades to explicate the relevance of a unitary conception, incorporating along the way consequential, not just evidentiary, considerations, which was critical to reasoning about the role of response format in educational measurement.

18.3 The Interplay of Constructs and Technology

The evolution of validity theory may have been essential to providing a compelling rationale for the use of constructed-response formats. However, the cost considerations in an educational setting are still an issue, especially in an educational context: According to Koretz and Hamilton (2006), "concerns about technical quality and costs are likely to dissuade most states from relying heavily on performance assessments in their accountability systems … particularly when states are facing heavy testing demands and severe budget constraints" (p. 536).

An important contribution by ETS to the development of constructed-response formats has been to take advantage of technological developments for educational and professional testing purposes. Among the most salient advances are the following:

- using computers to deploy constructed-response formats that expand construct coverage
- taking advantage of technology to enable more efficient human scoring
- pioneering research on automated scoring in a wide range of domains to improve cost effectiveness and further leverage the computer as a delivery medium

If the scanner enabled the large-scale use of multiple-choice tests, the advent of the computer played a similar role in enabling the large-scale use of constructed-response formats.[26] Incorporating technological advances into operational testing had been common practice at ETS almost from inception (Traxler 1951, 1954). However, a far more visionary perspective was apparent at the highest levels of the organization. In 1951, ETS officer William Turnbull coined the term, tailored testing (Lord 1980, p. 151); that is, the idea of adapting the test to the test taker.[27] Some years later, as the organization's executive vice president, he elaborated on it (Turnbull 1968):

[26] For perhaps the most complete history of the scanner and how it impacted testing, see Russell (2006, pp. 36–47).

[27] "Tailoring" a test was not a totally new idea, in the sense that Binet was practicing it at the turn of the twentieth century. Also, Cowden (1946) at Princeton University used sequential sampling, a method associated with quality control, as a test design (see Weiss and Betz 1973; Wood 1973). In

> The next step should be to provide examinations in which the individual questions are contingent on the student's responses to previous questions. If you will permit the computer to raise its ugly tapes, I would like to put forward the prospect of an examination in which, for each examinee, the sequence of questions is determined by his response to items earlier in the sequence. The questions will be selected to provide the individual student with the best opportunity to display his own profile of talent and accomplishment, without wasting time on tasks either well below or well beyond his level of developed ability along any one line. Looking farther down this same path, one can foresee a time when such *tailor-made tests* will be part and parcel of the school's instructional sequence; when the results will be accumulated and displayed regularly as a basis for instruction and guidance; and when the pertinent elements of the record will be banked as a basis for such major choice points as the student's selection of a college. (p. 1428, emphasis added)

Although Turnbull was not addressing the issue of format, his interest in computer-based testing is relevant to the eventual wider use of constructed-response formats, which perhaps would not have been feasible in the absence of computer-based testing. (The potential of microcomputers for testing purposes was recognized early at ETS; Ward 1984.) That an officer and future president of ETS would envision in such detail the use of computers in testing could have set the stage for an *earlier* use of computers for test delivery than might otherwise have been the case. And, if as Fowles (2012) argued, computer delivery was in part responsible for the adoption of writing in postsecondary admissions tests like the GRE General Test, then it is possible that the early adoption of computer delivery by ETS accelerated that process.[28] The transition to computer delivery started with what was later named the *ACCUPLACER*® test, a placement test consisting entirely of multiple-choice items developed for the College Board. It was first deployed in 1985 (Ward 1988). It is an important first success because it opened the door for other tests to follow.[29]

Once computer delivery was successfully implemented, it would be natural for other ETS programs to look into the possibility. Following the deployment of ACCUPLACER, the GRE General Test was introduced in 1992 (Mills and Steffen 2000). The 1992 examination was an adaptive test consisting of multiple-choice sections for Verbal Reasoning, Quantitative Reasoning, and Analytical Reasoning. However, the Analytical Reasoning measure was replaced in 2002 by the Analytical

---

fact, an experiment in the adaptive administration of the Stanford–Binet was reported as early as 1947 (Hutt 1947) and Hick (1951) shortly thereafter presented the essence of all the components of adaptive testing as we understand the term today.

[28] The contributions of Martha Stocking (1942–2006) in this process should be acknowledged. She was hired by Fred Lord in the 1960s and was soon making contributions to adaptive testing on her own (Stocking 1969). She made many contributions to adaptive testing over her career, especially in the area of controlling the exposure of individual test items (Stocking and Swanson 1993).

[29] Although it is entirely possible that while Turnbull may have been a visionary and could have encouraged Fred Lord to think about the idea of adaptive testing, Turnbull, apparently, was not involved in the decisions leading to the implementation of adaptive testing. ACCUPLACER (Ward 1988), the first ETS-produced adaptive test, was deployed in 1985 some years after Turnbull had resigned as president of ETS in 1981, according to Bill Ward (personal communication, July 6, 2010), who was the main developer of ACCUPLACER at ETS.

Writing section, consisting of two prompts: an issue prompt (45 minutes with a choice between two prompts) and an argument prompt (30 minutes).

The transition to computer delivery in 1992 and the addition of writing in 2002 appear to have flowed seamlessly, but in fact, the process was far more circuitous. The issue and argument prompts that composed the Analytical Writing measure were a significant innovation in assessment design and an interesting example of serendipity, the interplay of formats, technology, and attending to the consequences of testing.

Specifically, the design of the eventual GRE Analytical Writing measure evolved from the GRE Analytical Reasoning (multiple-choice) measure, which was itself a major innovation in the assessment of reasoning (Powers and Dwyer 2003). The Analytical Reasoning measure evolved by including and excluding different item types. In its last incarnation, it consisted of two multiple-choice item types, analytical reasoning and logical reasoning. The logical reasoning item type called for evaluating plausible conclusions, determining missing premises, finding the weakness of a conclusion, and so on (Powers and Dwyer 2003, p. 19). The analytical reasoning item type presented a set of facts and rules or restrictions. The test taker was asked to ascertain the relationships permissible among those facts, and to judge what was necessary or possible under the given constraints (Chalifour and Powers 1989).

Although an extensive program of research supported the development of the Analytical Reasoning measure, it also presented several challenges *especially under computer delivery*. In particular, performance on the logical reasoning items correlated highly with the verbal reasoning items, whereas performance on the analytical reasoning items correlated highly with quantitative reasoning items (Powers and Enright 1987), raising doubts about the construct it assessed. Moreover, no conclusive validity evidence for the measure as a whole was found when using an external criterion (Enright and Powers 1991). The ambiguous construct underpinnings of the Analytical Reasoning measure were compounded by the presence of speededness (Bridgeman and Cline 2004), which was especially harmful under computer delivery. Given the various challenges encountered by the Analytical Reasoning measure, it is no surprise that it ultimately was replaced by the Analytical Writing measure, which offered a well-balanced design.

The issue prompt has roots in the pedagogy of composition. As D'Angelo (1984) noted, textbooks dating back to the nineteenth century distinguish four genre: narration, description, exposition, and argumentation. Argumentation was defined as "the attempt to *persuade* others of the truth of a proposition" (p. 35, emphasis added). There is less precedent, if any, for the GRE argument prompt, which presents the task of *critiquing* an argument. The germ for the idea of an argument-critique prompt was planted during efforts to better prepare minority students for the GRE Analytical Reasoning measure, specifically, the logical reasoning item type (Peter Cooper, personal communication, November 27, 2013):

> The Logical Reasoning items … took the form of a brief stimulus passage and then one or more questions with stems such as "Which of the following, if true, weakens the argument?," "The argument above rests on which of the following assumptions," and so forth,

> with five options. At a workshop in Puerto Rico, a student commented that he would prefer questions that allowed him to comment on the argument in his own terms, not just pick an answer someone else formulated to a question someone else posed. I thought to myself, "Interesting concept … but be careful what you wish for" and did nothing for a couple of years, until [Graduate Management Admission Test] GMAT … told us in the summer of 1993 that it wanted to add a constructed-response measure, to be operational by October 1994, that would get at analytical reasoning—i.e., not just be another writing measure that rewarded fluency and command of language, although these would matter as well. Mary Fowles had discussed "Issue"-like prototypes with the [Graduate Management Admission Council] GMAC's writing advisory committee, which liked the item type but seemed to want something more "analytical" if possible. I recalled the student's comment and thought that a kind of constructed-response Logical Reasoning item could pair well with the Issue-type question to give a complementary approach to analytical writing assessment: In one exercise, students would make their own argument, developing a position on an issue, and in the other exercise they would critically evaluate the line of reasoning and use of evidence in an argument made by someone else. Both kinds of skills are important in graduate-level work.

Mary Fowles (2012) picked up the story from there: "What caused this seemingly rapid introduction of direct writing assessment for admission to graduate and professional programs?" (pp. 137–138). She cited factors such as the "growing awareness [of the relationship] between thinking and writing"; the availability of the computer as a delivery medium, which "enabled most examinees to write more fluently" and "streamlined the process of collecting written responses"; and "essay testing programs [that] now had the advantage of using automated scoring" (pp. 137–138).

Although the genesis of the argument prompt type came from attempts to help prepare students of diverse backgrounds for the multiple-choice GRE Analytical Reasoning section, the analytical writing measure comprising issue and argument prompts was used first by the GMAT. In 1994, that measure was offered in paper-and-pencil form, and then moved to computer when GMAT converted to an adaptive test in 1997. The GRE first used the measure as a stand-alone test (the GRE Writing Assessment) in 1999 and incorporated it into the General Test in 2002, as noted earlier.

The transition to computer delivery in the 1990s was not limited to the GRE and GMAT. The *TOEFL*® test transitioned as well. It evolved from a test conceived in the 1960s to a measure rooted in the communicative competence construct (Canale and Swain 1980; Duran et al. 1987). The earlier efforts to bolster TOEFL by introducing stand-alone writing and speaking tests—the Test of Written English (*TWE*® test) and the Test of Spoken English (*TSE*® test)—were seen as stopgap measures that led to an "awkward" situation for the "communication of score meaning" (C. A. Taylor and Angelis 2008, p. 37). Importantly, communicative competence called for evidence of proficiency in productive skills, which meant the assessment of writing *and* speaking proficiency in academic settings. In the case of speaking, these requirements meant that ultimately complex multimodal tasks were needed where students would read or listen to a stimulus and provide a spoken response. The construct of communicative competence was unpacked in frameworks corresponding to the four skills thought to compose it: reading (Enright et al. 2000), listening (Bejar

et al. 2000), writing (Cumming et al. 2000), and speaking (Butler et al. 2000). The frameworks served as the basis for experimentation, after which the blueprint for the test was set (Pearlman 2008a).

Computer delivery would prove critical to implementing such an ambitious test, especially the measurement of the productive skills. The inclusion of writing was relatively straightforward because there was already experience from GRE and GMAT. In fact, when it first transitioned to computer in 1998, TOEFL CBT used the TWE prompt as either a typed or handwritten essay. Nevertheless, there were still significant challenges, especially technological and assessment design challenges. The assessment of computer-delivered *speaking* on an international scale was unprecedented, especially considering the test security considerations.[30] The first generation of computer delivery that had served GRE and TOEFL CBT was less than ideal for effectively and securely delivering an international test administered every week. For one, testing that required speaking had the potential to interfere with other test takers. In addition, the quality of the speech captured needed to be high in all test centers to avoid potential construct-irrelevant variance. These requirements meant changes at the test centers, as well as research on the best microphones to capture spoken responses. On the back end, written and spoken responses needed to be scored quickly to comply with a turnaround of no more than 10 days. These requirements influenced the design of the next-generation test delivery system at ETS, iBT (Internet-based testing), and when the latest version of TOEFL was released in 2005, it was called the *TOEFL iBT®* test (Pearlman 2008a).

In addition to the technological challenges of delivering and scoring a secure speaking test, there were several assessment design challenges. To accommodate the international volume of test takers, it was necessary to administer the test 50 times a year. Clearly, the forms from week to week needed to be sufficiently different to prevent subsequent test takers from being able to predict the content of the test. The central concept was that of reusability, a key consideration in ECD, which was implemented by means of item templates (Pearlman 2008a).

18.3.1 Computer-Mediated Scoring

Once tests at ETS began to transition to computer delivery, computer-mediated scoring became of interest. Typically, faculty, in the case of educational tests, or practitioners, in the case of professional assessments, would congregate at a central location to conduct the scoring. As volume grew, best practices were developed, especially in writing (Baldwin 2004), and more generally (Baldwin et al. 2005; McClellan 2010). However, the increase in testing volumes called for better utilization of technology in the human scoring process.

[30] The IELTS assessment includes a speaking test but, unlike TOEFL, is administered locally by live examiners. Pearson currently offers a competitor to the TOEFL tests that includes writing and speaking measures that are scored solely by computer.

Perhaps anticipating the imminence of larger volumes, and the increasing availability of computers, there was experimentation with "remote scoring" fairly early[31] (Breland and Jones 1988). In the Breland and Jones study, the essays were distributed via courier to the raters at home. The goal was to evaluate whether solo scoring was feasible compared to centralized or conference scoring. Not surprisingly this form of remote scoring was not found to be as effective as conference scoring. All the affordances of computer technology were not taken advantage of until a few years later (Bejar and Whalen 2001; Driscoll et al. 1999; Kuntz et al. 2006).

The specialized needs of NAEP motivated a somewhat different use of the computer to mediate the human scoring process. In the early 1990s the NAEP program started to include state samples, which led to large increases in the volume of constructed responses. Such responses were contained in a single booklet for each student. To avoid potential scoring bias that would result from a single reader scoring all the constructed responses from a given student, a system was developed where the responses would be physically clipped and scanned separately. The raters would then score the scanned responses displayed on a terminal, with each response for a student routed to a different rater. The scoring of NAEP constructed responses was carried out by a subcontractor (initially NCS, and then Pearson after it acquired NCS) under direction from ETS. Scoring was centralized (all the raters were at the same location), but computer images of the work product were presented on the screen and the rater entered a score that went directly into a database.[32]

18.3.2 Automated Scoring

Though technology has had an impact on human scoring, a more ambitious idea was to automate the scoring of constructed responses. Page, a professor at the University of Connecticut, first proposed the idea for automated scoring of essays (Page 1966). It was an idea ahead of its time, because for automated scoring to be maximally useful, the responses need to be in digital form to begin with; digital test delivery was some decades away. However, as the computer began to be used for test delivery, even if it was limited to multiple-choice items, it was natural to study how the medium might be leveraged for constructed-response scoring purposes. Henry Braun, then vice president for research management, posed precisely that question (personal communication, July 9, 2014). Although a statistician by training, he was familiar with the literature on expert systems that had proliferated by the

[31] Even earlier, in the 1950s, Paul Diederich experimented enthusiastically with "lay readers," or "college-educated housewives" (Burke 1961, p. 258).

[32] Interestingly, during the same period, there was much activity at ETS related to scanning technology that had been developed for processing financial aid applications, led by Keith Reid-Green (1990). In fact, the seasonal nature of financial aid applications meant that the scanners ETS had could be used in support of other work, such as NAEP. However, in the end, the NAEP directors opted to use an external vendor.

1980s as a means of aiding and even automating expert judgment. In contrast to earlier research on actuarial judgment (Bejar et al. 2006), where the clinician and a regression equation were compared, in expert systems the role of the computer is more ambitious and consists of both analyzing an object (e.g., a doctor's course of treatment for a patient, an architectural design) and, based on that analysis, making a decision about the object, such as assigning a score level.

Randy Bennett took the lead at ETS in exploring the technology for scoring constructed responses in concert with theory about the relevant constructs, including mathematics (Bennett and Sebrechts 1996; Bennett et al. 1999, 2000a; Sandene et al. 2005; Sebrechts et al. 1991, 1996), computer science (Bennett and Wadkins 1995), graphical items (Bennett et al. 2000a; b), and formulating hypotheses (Bennett and Rock 1995). The scoring of mathematics items has reached a significant level of maturity (Fife 2013), as has the integration of task design and automated scoring (Graf and Fife 2012).

Much of the research on automated scoring was experimental, in the sense that actual applications needed to await the delivery of tests by computer. One ETS client, the National Council of Architectural Registration Boards (NCARB), was seriously considering on its own the implications of technology for the profession. The software used in engineering and architecture, computer-assisted design (CAD), was transitioning during the 1980s from minicomputers to desktop computers. A major implication of that transition was that the cost of the software came down significantly and became affordable to an increasingly larger number of architecture firms, thereby changing, to some extent, the entry requirements for the profession. Additionally, the Architectural Registration Examination introduced in 1983 was somewhat unwieldy, consisting of many parts that required several years to complete, since they could not all be taken together over the single testing window that was made available every June. A partnership between ETS and NCARB was established to transition the test to computer delivery and allow continuous testing, revise the content of the test, and take advantage of computer delivery, including automated scoring.

Management of the relationship between ETS and NCARB was housed in ETS's Center for Occupational and Professional Assessment (COPA), led by vice president Alice Irby, who was aware of the research on the utilization of computers for test delivery and scoring under Henry Braun. A project was initiated between ETS and NCARB that entailed developing new approaches to adaptive testing with multiple-choice items in a licensing context (Lewis and Sheehan 1990; Sheehan and Lewis 1992) and that had the more ambitious goal of delivering and scoring on computer the parts of the examination that required the demonstration of design skills.

The paper-and-pencil test used to elicit evidence of design skills included a very long design problem that took some candidates up to 14 hours to complete. Scoring such a work product was a challenge even for the practicing architects, called jurors. The undesirability of a test consisting of a single item from a psychometric perspective was not necessarily understood by the architects. However, they had realized that a single-item test could make it difficult for the candidate to recover from an

early wrong decision. That insight led to an assessment consisting of smaller constructed-response design tasks that required demonstrations of competence in several aspects of architectural practice (Bejar 2002; Bejar and Braun 1999). The process of narrowing the test design to smaller tasks was informed by practice analyses intended to identify the knowledge, skills, and abilities (so-called KSAs) required of architects, and their importance. This information was used to construct the final test blueprint, although many other considerations entered the decision, including considerations related to interface design and scorability (Bennett and Bejar 1998).

Reconceptualizing the examination to better comply with psychometric and technological considerations was a first step. The challenge of delivering and scoring the architectural designs remained. The interface and delivery, as well as supervising the engineering of the scoring engines, was led by Peter Brittingham, while the test development effort was led by Dick Devore. The scoring approach was conceived by Henry Braun (Braun et al. 2006) and Bejar (1991). Irv Katz contributed a cognitive perspective to the project (Katz et al. 1998). The work led to operational implementation in 1997, possibly the first high-stakes operational application of automated scoring.[33]

While ETS staff supported research on automated scoring in several domains, perhaps the ultimate target was essays, especially in light of their increasing use in high-volume testing programs. Research on automated scoring of textual responses began at ETS as part of an explicit effort to leverage the potential of technology for assessment. However, the first thorough evaluation of the feasibility of automated essay scoring was somewhat fortuitous and was carried out as a collaboration with an external partner. In the early 1990s, Nancy Petersen heard Ellis B. Page discuss his system, PEG, for scoring essays[34] at an AERA reception. Petersen suggested to Page the possibility of evaluating the system in a rigorous fashion using essays from 72 prompts taken from the *PRAXIS*® program, which had recently begun to collect essays on computer. The report (Page and Petersen 1995) was optimistic about the feasibility of automated scoring but lacked detail on the functioning of the scoring system. Based on the system's relatively positive performance, there was discussion between ETS and Page regarding a possible licensing of the system for nonoperational use, but the fact that Page would not fully reveal[35] the details of the system motivated ETS to invest further in its own development and research on automated scoring of essays. That research paid off relatively quickly since the system developed, the *e-rater*® engine, was put into operation in early 1999 to score GMAT

[33] The National Board of Medical Examiners (NBME) had also wanted to use automated scoring as part of its licensing test and had a project with that goal at about the same time the ETS and NCARB project was underway. Staff from both projects met informally over the years to exchange information. The NBME examination with automated scoring of Step 3 (Primum Computer Case Simulations) became operational in 2000 (P. Harik, personal communication, July 14, 2014), backed by a considerable body of research (Clauser 2000; Clauser et al. 2002; Clyman et al. 1995).

[34] The system is currently owned by Measurement Incorporated.

[35] According to Kaplan et al. (1995), the only feature that was revealed was essay length.

essays (Burstein et al. 1998). The system has continued to evolve (Attali and Burstein 2006; Burstein et al. 2004; Burstein et al. 2013) and has become a major ETS asset. Importantly, the inner workings of e-rater are well documented (Attali and Burstein 2006; Quinlan et al. 2009), and disclosed through patents.

The e-rater engine is an example of scoring based on linguistic analysis, which is a suitable approach for essays (Deane 2006). While the automated scoring of essays is a major accomplishment, many tests rely on shorter textual responses, and for that reason approaches to the scoring of short textual responses have also been researched. The basic problem of short-answer scoring is to account for the multiple ways in which a correct answer can be expressed. The scoring is then a matter of classifying a response, however expressed, into a score level. In the simplest case, the correct answer requires reference to a single concept, although in practice a response may require more than one concept. Full credit is given if all the concepts are present in the response, although partial credit is also possible if only some of the concepts are offered.

Whereas the score humans would assign to an essay can be predicted from linguistic features that act as correlates of writing quality, in the case of short responses, there are fewer correlates on which to base a prediction of a score. In a sense, the scoring of short responses requires an actual understanding of the *content* of the response so that it can be then be classified into a score level. The earliest report on short-answer scoring at ETS (Kaplan 1992) was an attempt to infer a "grammar" from a set of correct and incorrect responses that could be used to classify future responses. The approach was subsequently applied to scoring a computer-delivered version of a task requiring the generation of hypotheses (Kaplan and Bennett 1994). A more refined approach to short-answer scoring, relying on a more robust linguistic representation of responses, was proposed by Burstein et al. (1999), although it was not applied further.

As the complexities of scoring short answers became better understood, the complexity and sophistication of the approach to scoring grew as well. The next step in this evolution was the *c-rater*™ automated scoring engine (Leacock and Chodorow 2003).[36] The system was motivated by a need to lower the scoring load of teachers. Unlike earlier efforts, c-rater requires a *model* of the correct answer such that scoring a response is a matter of deciding whether it matches the model response. Developing such a model is not a simple task given the many equivalent ways of expressing the same idea. One of the innovations introduced by c-rater was to provide an interface to model the ideal response. In effect, a model response is defined by a set of possible paraphrases of the correct answer that are then represented in canonical or standard form. To evaluate whether a given response is in the set requires linguistic processing to deal with spelling and other issues so that the student response can be recast into the same canonical form as the model. The actual scoring is a matter of matching the student response against the model, guided by a set of linguistic rules. Because student responses can contain many spelling and

[36] The system was developed under an ETS subsidiary, ETS Technologies, which was ultimately folded back into R&D at ETS.

grammatical errors, the matching process is "fairly forgiving" (Leacock and Chodorow 2003, p. 396). The c-rater engine was evaluated in studies for NAEP (Sandene et al. 2005), and in other studies has been found useful for providing feedback to students (Attali 2010; Attali and Powers 2008, 2010). The most recent evaluation of the c-rater approach (Liu et al. 2014) took advantage of some refinements introduced by Sukkarieh and Bolge (2008). O. L. Liu et al. (2014) concluded that c-rater cannot replace human scores, although it has shown promise for use in low-stakes settings.

One limitation of c-rater is scalability. A scoring model needs to be developed for each question, a rather laborious process. A further limitation is that it is oriented to scoring responses that are verbal. However, short answers potentially contain numbers, equations, and even drawings.[37]

More recent approaches to short answer scoring have been developed including one referred to as Henry ML. Whereas c-rater makes an attempt to understand the response by identifying the presence of concepts, these newer approaches evaluate low-level aspects of the response, including "sparse features" like word and character n-grams, as well as "dense features" that compare the semantic similarity of a response to responses with agreed upon-scores (Liu et al. 2016; Sakaguchi et al. 2015).

The foregoing advances were followed by progress in the scoring of spoken responses. An automated approach had been developed during the 1990s by the Ordinate Corporation based on "low-entropy" tasks, such as reading a text aloud (Bernstein et al. 2000). The approach was, however, at odds with the communicative competence perspective that was by then driving the thinking of TOEFL developers. ETS experimented with automated scoring of high-entropy spoken responses (Zechner, Bejar, & Hemat, 2007). That is, instead of reading a text aloud, the tasks called for responses that were relatively extemporaneous and therefore more in line with a communicative perspective. The initial experimentation led rather quickly to an approach that could provide more comprehensive coverage of the speaking construct (Zechner et al. 2007b, 2009a). The current system, known as the *SpeechRater*℠ service, is used to score the *TOEFL Practice Online* (*TPO*™) test, which is modeled after the speaking component of the TOEFL. Efforts continue to further expand the construct coverage of the scoring engine by integrating additional aspects of speaking proficiency, such as content accuracy and discourse coherence (Evanini et al. 2013; Wang et al. 2013; Yoon et al. 2012). Additionally, the scope of applicability has been expanded beyond English as a second language (ESL) to also include the assessment of oral reading proficiency for younger students by means of low-entropy tasks (Zechner et al. 2009b, 2012). Importantly, the same underlying engine is used in this latter case, which argues well for the potential of that engine to support multiple types of assessments.

[37] The Smarter Balanced consortium, for example, field tested in 2014 such item types (Smarter Balanced Assessment Consortium 2014).

18.3.3 Construct Theory and Task Design

Technology was as important to the adoption of constructed-response formats as it was for the multiple-choice format, where the scanner made it possible to score large volumes of answer sheets. However, much more was needed in the case of constructed-response formats besides technology. Invariably, progress was preceded or accompanied by work on construct definition.

18.3.3.1 Writing

The publication that may have been responsible for the acceptance of holistic scoring (Godshalk et al. 1966) was, in fact, an attempt to empirically define the writing construct. Over the years, many other efforts followed, with various emphases (Breland 1983; Breland and Hart 1994; Breland et al. 1984, 1987). Surveys of graduate faculty identified written argumentation, both constructing and critiquing arguments, as an important skill for success in graduate school (Enright and Gitomer 1989). Summaries of research through 1999 (Breland et al. 1999) show convergence on various issues, especially the importance of defining the construct, and then designing the test accordingly to cover the intended construct, while simultaneously avoiding construct-irrelevant variance. In the case of the GMAT and GRE,[38] a design consisting of two prompts, creating and evaluating arguments, emerged after several rounds of research (Powers et al. 1999a). The design remains in GMAT and GRE.

Writing was partially incorporated into the TOEFL during the 1980s in the form of the TWE. It was a single-prompt "test." A history of the test is provided by Stansfield (1986a). With plans to include writing in the revised TOEFL, more systematic research among English language learners began to emerge, informed by appropriate theory (Hamp-Lyons and Kroll 1997). Whereas the distinction between issue and argument is thought to be appropriate for GRE and GMAT, in the case of TOEFL the broader construct of communicative competence has become the foundation for the test. With respect to writing, a distinction is made between an independent and an integrated prompt. The latter requires the test takers to refer to a document they read as part of the prompt. (See TOEFL 2011, for a brief history of the TOEFL program.)

Understandably, much of the construct work on writing has emphasized the postsecondary admissions context. However, in recent years, K-12 education reform efforts have increasingly incorporated test-based accountability approaches (Koretz and Hamilton 2006). As a result, there has been much reflection about the nature of school-based testing. The research initiative known as CBAL (Cognitively Based Assessment *of*, *for*, and *as* Learning) serves as an umbrella for experimentation on

[38] Today, the GMAT is administered and developed under the auspices of the Graduate Management Admissions Council (GMAC). It was originally developed at ETS for the GMAC and shared item types and staff with the GRE program. The interest in incorporating writing in the GMAT dates back to at least the mid-1980s (Owens 2006).

next-generation K–12 assessments. Under this umbrella, the writing construct has expanded to acknowledge the importance of other skills, specifically reading and critical thinking, and the developmental trajectories that underlie proficiency (Deane 2012; Deane and Quinlan 2010; Deane et al. 2008, 2012). In addition to expanding the breadth of the writing construct, recent work has also emphasized depth by detailing the nature of the evidence to be sought in student writing, especially argumentative writing (Song et al. 2014). Concomitant advances that would enable automated scoring for rich writing tasks have also been put forth (Deane 2013b).

18.3.3.2 Speaking

The assessment of speaking skills has traditionally taken place within an ESL context. The TSE (Clark and Swinton 1980) was the first major test of English speaking proficiency developed at ETS. Nevertheless, Powers (1984) noted that among the challenges facing the development of speaking measures were construct definition and cost. With respect to construct definition, a major conference was held at ETS in the 1980s (Stansfield 1986b) to discuss the relevance of communicative competence for the TOEFL. Envisioning TOEFL from that perspective was a likely outcome of the conference (Duran et al. 1987). Evidence of the acceptance of the communicative competence construct can be seen in its use to validate TSE scores (Powers et al. 1999b), and in the framework for incorporating a speaking component in a revised TOEFL (Butler et al. 2000). The first step in the development of an operational computer-based speaking test was the TOEFL Academic Speaking Test (TAST), a computer-based test intended to familiarize TOEFL test takers with the new format. TAST was introduced in 2002 and served to refine the eventual speaking measure included in TOEFL iBT. Automated scoring of speaking as discussed above, could help to reduce costs, but is not yet sufficiently well developed (Bridgeman et al. 2012). The *TOEIC*® Speaking and Writing test followed the TOEFL (Pearlman 2008b) in using ECD for assessment design (Hines 2010) as well as in the inclusion of speaking (Powers 2010; Powers et al. 2009).

18.3.3.3 Mathematics

Constructed-response items have been standard in the AP program since inception and were already used in NAEP by 1990 (Braswell and Kupin 1993). The SAT relied on multiple-choice items for much of its history (Lawrence et al. 2002) but also introduced in the 1990s a simple constructed-response format, the grid-in item, that allowed students to enter numeric responses. Because of the relative simplicity of numeric responses, they could be recorded on a scannable answer sheet, and therefore scored along with the multiple-choice responses. Various construct-related considerations motivated the introduction of the grid-in format, among them the influence of the standards produced by the National Council of Teachers of

Mathematics (Braswell 1992) but also considerations about the response process. For example, Bridgeman (1992) argued that in a mathematics context, the multiple-choice format could provide the student inadvertent hints and also make it possible to arrive at the right answers by reasoning backward from the options. He evaluated the SAT grid-in format with GRE items and concluded that the multiple-choice and grid-in versions of GRE items behaved very similarly. Following the adoption of the grid-in format in the SAT, a more comprehensive examination of mathematics item formats that could serve to elicit quantitative skills was undertaken, informed by advances in the understanding of mathematical cognition and a maturing computer-based infrastructure (Bennett and Sebrechts 1997; Bennett et al. 1997, 1999, 2000a, b; Sandene et al. 2005; Sebrechts et al. 1996). More recently, the mathematics strand of the CBAL initiative has attempted to unpack mathematical proficiency by means of competency models, the corresponding constructed-response tasks (Graf 2009), and scoring approaches (Fife 2013).

18.3.3.4 History[39]

A design innovation introduced by the AP history examinations was the document-based question (DBQ). Such questions require the test taker to incorporate, in a written response, information from one or more historical documents.[40] The idea for the format was based on input from a committee member who had visited libraries in England and saw that there were portfolios of primary historical documents, which apparently led to the DBQ. The DBQ was first used with the U.S. History examination, and the European History examination adopted the format the following year, as did World History when it was introduced in 2002. The scoring of document-based responses proved to be a challenge initially, but since its rationale was so linked to the construct, the task has remained.

18.3.3.5 Interpersonal Competence

Interpersonal competence has been identified as a twenty-first-century educational skill (Koenig 2011) as well as a workforce skill (Lievens and Sackett 2012). The skill was assessed early on at ETS by Larry Stricker (Stricker 1982; Stricker and Rock 1990) in a constructed-response format by means of videotaped stimuli, a relatively recent invention at the time. The recognition of the affordances of

[39] This section is based on an interview conducted on May 12, 2011, with Despina Danos, a senior Advanced Placement assessment developer.

[40] Although it is safe to say that assessments in the 1960s did not flow from a comprehensive framework or the explication of the target constructs, the current construct statement for AP History includes the skill of "crafting historical arguments from historical evidence" (College Board 2011, p. 8).

technology appears to have been the motivation for the work (Stricker 1982): "The advent of videotape technology raises new possibilities for assessing interpersonal competence because videotape provides a means of portraying social situations in a comprehensive, standardized, and economical manner" (p. 69).

18.3.3.6 Professional Assessments

Historically, ETS tests have been concerned with aiding the transition to the next educational level and, to a lesser extent, with tests designed to certify professional knowledge. Perhaps the earliest instance of this latter line of work is the "in-basket test" developed by Frederiksen et al. (1957). Essentially, the in-basket format is used to simulate an office environment where the test taker plays the role of school principal or business executive, for example. The format was used in an extended study concerned with measurement of the administrative skills of school principals in a simulated school (Hemphill et al. 1962). Apart from the innovative constructed-response format, the assessment was developed following what, in retrospect, was a very sophisticated assessment design approach. First, a job analysis was conducted to identify the skills required of an elementary school principal. In addition, the types of problems an elementary school principal is confronted with were identified and reduced to a series of incidents. This led to a universe of potential items by combining the problems typically confronted with the skills assumed to be required to perform as a principal based on the best research at the time (Hemphill et al. 1962, p. 47). Three skills were assumed to be (a) technical, (b) human, and (c) conceptual. The four facets of the jobs were taken to be (a) improving educational opportunity, (b) obtaining and developing personnel, (c) maintaining effective interrelationships with the community, and (d) providing and maintaining funds and facilities. The crossing of skill and facets led to a 4 × 3 matrix. Items were then written for each cell.

While the research on the assessment of school principals was highly innovative, ETS also supported the assessment of school personnel with more traditional measures. The first such assessment was bequeathed to the organization when the American Council on Education transferred the National Teacher Examination (NTE) in 1948 to the newly founded ETS. However, in the early 1980s, under President Greg Anrig[41] a major rethinking of teacher testing took place and culminated in the launching, in 1993, of the *PRAXIS SERIES*® tests. The *PRAXIS I*® and *PRAXIS II*® tests were concerned with content and pedagogical knowledge measured by multiple-choice items, as well as some types of constructed-response tasks. However, the *PRAXIS III*® tests were concerned with classroom performance and involved observing teachers in situ, a rather sharp departure from traditional mea-

[41] Greg Anrig was ETS's third president from 1981 until his death in 1993.

surement approaches. Although classroom observation has long been used in education, PRAXIS III appears to be among the first attempts to use *observations-as-measurement* in a classroom context. The knowledge base for the assessment was developed over several years (Dwyer 1994) and included scoring rubrics and examples of the behavior that would be evidence of the different skills required of teachers. The PRAXIS III work led to the Danielson *Framework for Teaching*,[42] which has served as the foundation for school-leader evaluations of teachers in many school districts, as well as for video-based products concerned with evaluation,[43] including those of the MET project (Bill and Melinda Gates Foundation 2013).

Whereas PRAXIS III was oriented toward assessing beginning teachers, ETS was also involved with the assessment of master teachers as part of a joint project with the National Board of Professional Teaching Standards (NBPTS). The goal of the assessment was to certify the expertise of highly accomplished practitioners. Pearlman (2008a, p. 88) described the rich history as "a remarkable journey of design, development, and response to empirical evidence from practice and use," including the scoring of complex artifacts. Gitomer (2007) reviewed research in support of NBPTS.

COPA was devoted to developing assessments for licensing and certification in fields outside education. In addition to that of architects mentioned earlier, COPA also considered the licensing of dental hygienists (Cameron et al. 2000; Mislevy et al. 1999, 2002b), which was one of the earliest applications of the ECD framework that will be discussed next.

18.3.3.7 Advances in Assessment Design Theory

For most of the twentieth century, there did not exist a *comprehensive* assessment design framework that could be used to help manage the complexity of developing assessments that go beyond the multiple-choice format. Perhaps this was not a problem because such assessments were relatively few and any initial design flaws could be remedied over time. However, several factors motivated the use of more ambitious designs, including the rapid technological innovations introduced during the second half of the twentieth century, concerns about the levels of achievement and competitiveness of U.S. students, the continued interest in forms of assessment beyond the multiple-choice item, and educational reform movements that have emphasized test-based accountability. A systematic approach to the design of complex assessments was needed, including ones involving the use of complex constructed responses.

[42] http://danielsongroup.org/framework/

[43] http://www.teachscape.com/

ECD is rooted in validity theory. Its genesis (Mislevy et al. 2006) is in the following quote from Messick (1994) concerning assessment design, which, he argued,

> would begin by asking what complex of knowledge, skills, and other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, *what behaviors or performances should reveal those constructs*, and *what task or situations should elicit those behaviors*? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17, emphasis added)

ECD is a fleshing out of the quote into a comprehensive framework consisting of interlocking models. The *student model* focuses on describing the test taker, whereas the *evidence model* focuses on the nature and analysis of the responses. The evidence model passes its information to the student model to update the characterization of what the examinee knows and can do. Finally, the *task model* describes the items. Thus, if the goal is to characterize the students' communicative competence, an analysis of the construct is likely to identify writing and speaking skills as components, which means the student model should include characterizations of these student skills. With that information in hand, the details of the evidence model can be fleshed out: What sort of student writing and speaking performance or behavior constitutes evidence of students' writing and speaking skills? The answer to that question informs the task models, that is, what sorts of tasks are required to elicit the necessary evidence? ECD is especially useful in the design of assessments that call for constructed responses by requiring the behavior that constitutes relevant evidence of writing and speaking skills, for example, to be detailed and then prescribing the task attributes that would elicit that behavior. The evidence model, apart from informing the design of the tasks, is also the basis for scoring the responses (Mislevy et al. 2006).

ECD did not become quickly institutionalized at ETS, as Zieky (2014) noted. Nevertheless, over time, the approach has become widely used. Its applications include science (Riconscente et al. 2005), language (Mislevy and Yin 2012), professional measurement (Mislevy et al. 1999), technical skills (Rupp et al. 2012), automated scoring (Williamson et al. 2006), accessibility (Hansen and Mislevy 2008; T. Zhang et al. 2010), and task design and generation (Huff et al. 2012; Mislevy et al. 2002a). It has also been used to different degrees in the latest revisions of several ETS tests, such as TOEFL (Pearlman 2008b), in revisions of the College Board's AP tests (Huff and Plake 2010), and by the assessment community more generally (Schmeiser and Welch 2006, p. 313). Importantly, ECD is a broad design methodology that is not limited to items as the means of eliciting evidence. Games and simulations are being used with increasing frequency in an educational context, and ECD is equally applicable in both cases (Mislevy 2013; Mislevy et al. 2014, 2016).

18.3.4 Conclusion

It is clear that at ETS the transition to computer-based test delivery began early on and was sustained. The transition had an impact on constructed responses by enabling their use earlier than might have otherwise been the case. As Table 18.1 shows, online essay scoring and automated essay scoring were part of the transition for three major admissions tests: GMAT, GRE and TOEFL.

18.4 School-Based Testing

Although postsecondary admissions tests have been the main form of operational testing at ETS, school-based testing has been and continues to be an important focus. The Sequential Tests of Educational Progress (STEP) was an early ETS product in this domain. At one point it included a writing test that consisted of multiple-choice questions and an essay,[44] although it is no longer extant. By contrast, ETS involvement in two major twentieth-century school-based assessments, the *Advanced Placement Program*® examinations and the NAEP assessments, as well as in state assessments has grown. Constructed-response formats have played a major role, especially in AP and NAEP. In addition, the CBAL initiative has been prominent in recent years. They are discussed further in this section.

18.4.1 Advanced Placement

While the use of constructed responses encountered resistance at ETS in the context of admissions testing, the same was not true for the AP program, introduced in the mid-1950s. From the start, the AP program was oriented to academically advanced students who would be going to college, specifically to grant college credit or advanced placement by taking an examination. The seeds for the program were two reports (Lacy 2010), one commissioned by Harvard president James Bryant Conant (Committee on the Objectives of a General Education in a Free Society 1945), the other (General Education in School and College 1952) also produced at Harvard. These reports led to a trial of the idea in an experiment known as the Kenyon Plan.[45]

The eventual acquisition of the program by the College Board was not a given. Valentine (1987) noted that "Bowles [College Board president at the time] was not

[44] For a review, see Croon Davis et al. (1959).

[45] ETS was involved in the development and scoring of the Kenyon Plan before College Board agreed to take the program (Valentine 1987, p. 84).

Table 18.1 Writing assessment milestones for GMAT, GRE and TOEFL tests

Milestone	GMAT	GRE	TOEFL
When was writing introduced?	GMAT Analytic Writing Assessment (AWA) was introduced in the paper testing program by ETS in October 1994. The test consisted of one issue and one argument prompt.	GRE introduced the stand-alone GRE Writing Assessment in 1999, which consisted of one issue and one argument prompt (students were given a choice between two issue prompts). In 2002 the Analytical Writing measure replaced the Analytical reasoning section and became part of the GRE General Test. In 2011, the choice of issue prompts was removed and prompts variants were introduced.	Test of Written English portion of the paper-based TOEFL test was introduced in 1986 at selected administrations of TOEFL. The 1998 TOEFL CBT writing task consisted of a choice of handwritten or keyed essay, essentially the same task as the Test of Written English portion of the paper-based TOEFL test. A writing measure consisting of integrated and independent prompts was introduced with the release of TOEFL iBT in 2006.
When was computer-based/adaptive testing introduced?	GMAT switched entirely from paper-and-pencil to on-demand CAT in October 1997.	GRE switched to on-demand CAT in 1992 and abandoned CAT in favor of MST in 2011.	TOEFL CBT on-demand testing was introduced in 1998. The CBT essay score was combined with the Structure selected-response subsection to report out on a Structure Writing section score. Listening and Structure were adaptive.
When was online scoring deployed?	Under on-demand testing, scores need to be reported on an ongoing basis and that, in turn, requires continuous scoring. The Online Scoring Network (OSN) was developed for that purpose and was first used operationally in October 1997 for GMAT essays.	OSN has been used to score GRE essays since 1999 when the GRE Writing Assessment was introduced.	Online scoring was deployed when TOEFL CBT was launched in 1998.

(continued)

Table 18.1 (continued)

Milestone	GMAT	GRE	TOEFL
When was automated scoring introduced?	Automated scoring with e-rater as a contributory score started in January 1999.	e-rater scoring as a check score was introduced in 2008.	e-rater started as contributory score to independent writing for TOEFL iBT beginning July 2009; contributory score for integrated writing began November 2010.

Note. *CAT* = computer adaptive testing, *CBT* = computer based testing, *MST* = multistage testing

sure that taking the program was in the Board's interest" (p. 85). Initially, some of the AP examinations were entirely based on constructed responses, although eventually all, with the exception of Studio Art, included a mix of constructed-response and multiple-choice items. A program publication, *An Informal History of the AP Readings 1956–1976* (Advanced Placement Program of the College Board 1980), provides a description of the scoring process early in the program's history.

Interestingly, in light of the ascendancy of the multiple-choice format during the twentieth century, the use of constructed responses in AP does not appear to have been questioned. Henry Dyer, a former ETS vice president (1954–1972), seems to have been influential in determining the specifications of the test (Advanced Placement Program of the College Board 1980, p. 2). Whereas Dyer did not seem to have been opposed to the use of constructed responses in the AP program, he was far more skeptical of their value in the context of another examination being conceived at about the same time, the Test of Developed Ability.[46] In discussing the creation of that test, Dyer (1954) noted that

> there may be one or two important abilities which are measureable only through some type of free response question. If an examining committee regards such abilities as absolutely vital in its area, it should attempt to work out one or two free response questions to measure them. Later on, we shall use the data from the tryouts to determine whether the multiple-choice sections of the test do not in fact measure approximately the same abilities as the free

[46] The Test of Developed Ability is a nearly forgotten test. It is relevant to this chapter since in its original conception, it employed constructed responses. Henry Chauncey was a champion for the test at ETS (Lemann 1999, p. 95). According to N. Elliot (2005, p. 149), citing Henry Dyer, Frank Bowles, president of College Board, proposed the test as early as 1949. Bowles thought that there were "changes coming in the kind of tests that would be suitable for college admission." The Test of Developed Ability was designed to measure achievement, in contrast to the SAT, which was oriented, at the time, toward measuring ability. The design of the Test of Developed Ability called for constructed responses, which presented a major scoring hurdle and may have been one of the reasons the test never became operational. According to Lemann (1999), the projected cost of the test was six dollars, as opposed to three dollars for the SAT. This work transpired during the 1950s when some individuals thought there should be an alternative to the SAT that was more achievement oriented. In fact, such an alternative led to the founding of ACT in 1959, led by Lindquist (see, N. Elliot 2014, p.246). For further discussion of the Test of Developed Ability, see N. Elliot (2005, p. 148; 2014, p.292) and Lemann (1999).

> response sections. If they do, the free response section will be dropped, if not, they will be retained. (p. 7)

Thus, there was realization that the AP program was unique with respect to other tests, in part because of its use of constructed responses. In *An Informal History of the AP Readings 1956–76* (Advanced Placement Program of the College Board 1980), it was noted that

> neither the setting nor the writing of essay examination was an innovation. The ancient Chinese reputedly required stringent written examinations for high government offices 2,500 years ago. European students have long faced pass-or-perish examinations at the end of their courses in the Lycée, Gymnasium, or British Secondary system. In this country, from 1901 to 1925, the College Board Comprehensives helped to determine who would go to the best colleges. But the Advanced Placement Program was new, and in many ways unique. (p. 2)

As the College Board's developer and administrator for the AP program, ETS has conducted much research to support it. The contributions focused on fairness (e.g., Breland et al. 1994; Bridgeman et al. 1997; Dorans et al. 2003; Stricker and Ward 2004), scoring (e.g., Braun 1988; Burstein et al. 1997; Coffman and Kurfman 1968; Myford and Mislevy 1995; Zhang et al. 2003), psychometrics (e. g., Bridgeman et al. 1996a, b; Coffman and Kurfman 1966; Lukhele et al. 1994; Moses et al. 2007), and validity and construct considerations (e. g., Bennett et al. 1991; Bridgeman 1989; Bridgeman and Lewis 1994).

18.4.2 Educational Surveys[47]

As noted earlier, NAEP has been a locus of constructed-response innovation at ETS. NAEP was managed by the Education Commission of the States until 1983 when ETS was awarded the contract to operate it. With the arrival of NAEP, ETS instituted matrix sampling, along with IRT (Messick et al. 1983); both had been under development at ETS under Fred Lord,[48] and both served to undergird a new approach to providing the "Nation's Report Card" in several subjects, with extensive use of constructed-response formats. To NAEP's credit, explicating the domain of knowledge to be assessed by means of "frameworks" had been part of the assessment development process from inception. Applebee (2007) traced the writing framework back to 1969. Even before that date, however, formal frameworks pro-

[47] For a fuller discussion of educational surveys see Beaton and Barone (Chap. 8, this volume) and Kirsch et al. (Chap. 9, this volume).

[48] Lord (1965) credits William Turnbull with the essence of the idea of item sampling and Robert L. Ebel with its application to norming. An implication of item and matrix sampling for constructed-response formats is that they make it possible to administer a large number of items, without any one student responding to a long test, by assigning subsets of items to different students. The idea can be leveraged for school-based testing (Bejar and Graf 2010; Bock and Mislevy 1988).

viding the rationale for content development were well documented (Finley and Berdie 1970). The science framework refers early on to "inquiry skills necessary to solve problems in science, specifically the ability to recognize scientific hypotheses" (p. 14). The assessment of inquiry skills has since become standard in science assessment but was only recently implemented operationally with the redesigned AP exams.

NAEP has been the source of multiple content and psychometric innovations (Mazzeo et al. 2006), including the introduction of mixed-format assessment consisting of both multiple choice and the large-scale use of constructed-response items. Practical polytomous IRT was developed in a NAEP context as documented by Carlson and von Davier (Chap. 5, this volume), and, as described earlier, NAEP introduced innovations concerned with the scoring of written responses. Finally, ETS continues to collaborate with NAEP in the exploration of technological advances to testing (Bennett et al. 2010).[49] The transition to digital delivery is underway as of this writing. In fact, the 2017 writing assessment was administered on tablets supplied by NAEP and research into the use of mixed-format adaptive in mathematics has also been carried out (Oranje et al. 2014).

18.4.3 Accountability Testing

The start of K–12 testing in the United States dates back to the nineteenth century, when Horace Mann, an educational visionary, introduced several innovations into school testing, among them the use of standardized (written constructed-response) tests (U.S. Congress and Office of Technology Assessment 1992, chapter 4). The innovations Mann introduced were, in part, motivated by a perception that schools were not performing as well as could be expected. Such perceptions have endured and have continued to fuel the debate about the appropriate use of tests in K–12. More recently, the Nation at Risk report (National Commission on Excellence in Education 1983) warned that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people" (para. 1). Similarly, the linking of the state of education to the nation's economic survival[50] was behind one effort in the early 1990s (U.S. Department of Labor and Secretary's Commission on Achieving Necessary Skills 1991; known as SCANS), and it had significant implications for the future of testing. As Linn (1996) noted, the system of assessment expected to emerge from the SCANS effort and to be linked to instruction, "would require direct appraisals of student performance" (p. 252) and would serve to promote the measured skills.

[49] NAEP has set up a website on technology-based assessments: http://nces.ed.gov/nationsreportcard/tba/

[50] The concerns continue and have been described as a "perfect storm" (Kirsch et al. 2007).

The calls for direct assessment that promotes learning joined the earlier Frederiksen (1984) assault on the multiple-choice item type, which had been heard loudly and clearly, judging by the number of citations to that article.[51] The idea of authentic assessment (Wiggins 1989) as an alternative to the standardized multiple-choice test, took hold among many educators, and several states launched major performance-based assessments. ETS participated in exploring these alternatives, especially portfolio assessment (Camp 1985, 1993), including in their evaluation in at least one state, California (Thomas et al. 1998).

Stetcher (2010) provided a detailed review of the different state experiments in the early 1990s in Vermont, Kentucky, Maryland, Washington, California, and Connecticut. A summary of a conference (National Research Council 2010, p. 36) noted several factors that led to the demise of these innovative programs:

- Hurried implementation made it difficult to address scoring, reliability, and other issues.
- The scientific foundation required by these innovative assessments was lacking.
- The cost and burden to the school was great, and questions were raised as to whether they were worth it.
- There were significant political considerations, including cost, time, feasibility of implementation, and conflicts in purpose among constituencies.

Not surprisingly, following this period of innovation, there was a return to the multiple-choice format. Under the No Child Left Behind (NCLB) legislation,[52] the extent of federally mandated testing increased dramatically and once again the negative consequences of the predominant use of multiple-choice formats were raised. In response, a research initiative was launched at ETS known as CBAL (Bennett and Gitomer 2009).[53] Referring to the circumstances surrounding accountability testing under NCLB, Bennett and Gitomer noted,

> In the United States, the problem is … an accountability assessment system with at least two salient characteristics. The first characteristic is that there are now significant consequences for students, teachers, school administrators, and policy makers. The second characteristic is, paradoxically, very limited educational value. This limited value stems from the fact that our accountability assessments typically reflect a shallow view of proficiency defined in terms of the skills needed to succeed on relatively short and, too often, quite artificial test items (i.e., with little direct connection to real-world contexts). (p. 45)

The challenges that needed to be overcome to develop tests based on a deeper view of student achievement were significant and included the fact that more meaningful tests would require constructed-response formats to a larger degree, which required a means of handling the trade-off between reliability and time. As Linn and Burton (1994), and many others, have reminded us regarding constructed-response tests, "a substantial number of tasks will still be needed to have any reasonable level of confidence in making a decision that an individual student has or has not met the

[51] As of August 2014, it had been cited 639 times.

[52] http://www.ed.gov/policy/elsec/leg/esea02/index.html

[53] A website describing the CBAL initiative can be found at https://www.ets.org/cbal

standard" (p. 10). Such a test could not reasonably be administered in a single seating. A system was needed in which tests would be administered at more than one occasion. A multi-occasion testing system raises methodological problems of its own, as was illustrated by the California CLAS assessment (Cronbach et al. 1995). Apart from methodological constraints, increasing testing time could be resented, unless the tests departed from the traditional mold and actually promoted, not just probed, learning. This meant that the new assessment needed to be an integral part of the educational process. To help achieve that goal, a theory of action was formulated (Bennett 2010) to link the attributes of the envisioned assessment system to a set of hypothesized action mechanisms leading to improved student learning. (Of course, a theory of action *is* a theory, and whether the theory is valid is an empirical question.)

Even with a vision of an assessment system, and a rationale for how such a vision would lead to improved student learning, considerable effort is required to explicate the system and to leverage technology to make such assessments scalable and affordable. The process entailed the formulation of competency models for specific domains, including reading (Sheehan and O'Reilly 2011), writing (Deane et al. 2012), mathematics (Graf 2009), and science (Liu et al. 2013); the elaboration of constructs, especially writing (Song et al. 2014); and innovations in automated scoring (Deane 2013a, b; Fife 2013) and task design (Bennett 2011; Sheehan and O'Reilly 2011).

The timing of the CBAL system coincided roughly with the start of a new administration in Washington that had educational plans of its own, ultimately cast as the Race to the Top initiative.[54] The assessments developed under one portion of the Race to the Top initiative illustrate a trend toward the use of significant numbers of items requiring constructed responses. In addition, technology is being used more extensively, including adaptive testing by the Smarter Balanced Assessment Consortium, and automated scoring by some of its member states.

18.4.4 Conclusion

Admissions testing has been the primary business at ETS for most of its existence. Constructed-response formats were resisted for a long time in that context, although in the end they were incorporated. By contrast, the same resistance was not encountered in some school assessments, where they were used from the start in the AP program as well as in the NAEP program. The CBAL initiative has continued and significantly expanded that tradition by conceiving of instructionally rich computer-based tasks grounded in scientific knowledge about student learning.

[54] The tone for the initiative was oriented to increased reliance on constructed-response formats, as noted by President Obama: "And I'm calling on our nation's governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test" (White House 2009).

18.5 Validity and Psychometric Research Related to Constructed-Response Formats

The foregoing efforts occurred in the context of a vigorous validity and psychometric research program over several decades in support of constructed-response formats. It is beyond the scope of this chapter to review the literature resulting from that effort. However, the scope of the research is noteworthy and is briefly and selectively summarized below.

18.5.1 Construct Equivalence

The choice between multiple-choice or constructed-response format, or a mix of the two, is an important design question that is informed by whether the two formats function in similar ways. The topic has been approached conceptually and empirically (Bennett et al. 1990, 1991; Bridgeman 1992; Enright et al. 1998; Katz et al. 2000; Messick 1993; Wainer and Thissen 1993; Ward 1982; Ward et al. 1980).

18.5.2 Predictive Validity of Human and Computer Scoring

The predictive validity of tests based on constructed responses scored by humans and computers has not been studied extensively. A study (Powers et al. 2002) appears to be one of the few on the subject. More recently, Bridgeman (2016) showed the impressive psychometric predictive power of the GRE and TOEFL writing assessments.

18.5.3 Equivalence Across Populations and Differential Item Functioning

The potential incomparability of the evidence elicited by different test formats has fairness implications and not surprisingly has received much attention (e.g., Breland et al. 1994; Bridgeman and Rock 1993; Dorans 2004; Dorans and Schmitt 1993; Schmitt et al. 1993; Zwick et al. 1993, 1997). The challenges of differential item functioning across language groups have also been addressed (Xi 2010). Similarly, the role of different response formats when predicting external criterion measures has been investigated (Bridgeman and Lewis 1994), as have the broader implications of format for the admissions process (Bridgeman and McHale 1996).

18.5.4 Equating and Comparability

The use of constructed-response formats presents many operational challenges. For example, ensuring the comparability of scores from different forms is equally applicable to tests comprising constructed-response items as it is for multiple-choice tests. The primary approach to ensuring score comparability is through equating (Dorans et al. 2007), a methodology that had been developed for multiple-choice tests. As the use of constructed-response formats has grown, there has been an increase in research concerning equating of tests composed entirely, or partly, of constructed responses (Kim and Lee 2006; Kim and Walker 2012; Kim et al. 2010). Approaches to achieving comparability without equating, which rely instead on designing *tasks* to be comparable, have also been studied (Bejar 2002; Bridgeman et al. 2011; Golub-Smith et al. 1993).

18.5.5 Medium Effects

Under computer delivery, task presentation and the recording of responses is very different for multiple-choice and constructed-response items. These differences could introduce construct-irrelevant variance due to the testing medium. The investigation of that question has received significant attention (Gallagher et al. 2002; Horkay et al. 2006; Mazzeo and Harvey 1988; Powers et al. 1994; Puhan et al. 2007; Wolfe et al. 1993).

18.5.6 Choice

Students' backgrounds can influence their interest and familiarity with the topics presented in some types of constructed-response items, which can lead to an unfair assessment. The problem can be compounded by the fact that relatively few constructed-response questions can be typically included in a test since responding to them is more time consuming. A potential solution is to let students choose from a set of possible questions rather that assigning the same questions to everyone. The effects of choice have been investigated primarily in writing (Allen et al. 2005; Bridgeman et al. 1997; Lukhele et al. 1994) but also in other domains (Powers and Bennett 1999).

18.5.7 Difficulty Modeling

The difficulty of constructed-response items and the basis for, and control of, variability in difficulty have been studied in multiple domains, including mathematics (Katz et al. 2000), architecture (Bejar 2002), and writing (Bridgeman et al. 2011; Joe et al. 2012).

18.5.8 Diagnostic and Formative Assessment

Diagnostic assessment is a broad topic that has much in common with formative assessment because in both cases it is expected that the provided information will lead to actions that will enhance student learning. ETS contributions in this area have included the development of psychometric models to support diagnostic measurement based on constructed responses. Two such developments attempt to provide a psychometric foundation for diagnostic assessments. Although these efforts are not explicitly concerned with constructed responses, they support such assessments by accommodating polytomous responses. One approach is based on Bayesian networks (Almond et al. 2007), whereas the second approach follows a latent variable tradition (von Davier 2013).

18.6 Summary and Reflections

The multiple-choice item format is an early-twentieth-century American invention. Once the format became popular following its use in the Army Alpha and SAT, it became difficult for constructed-response formats to regain a foothold. The psychometric theory that also emerged in the early twentieth century emphasized score reliability and predictive validity. Those emphases presented further hurdles. The interest in constructed-response formats, especially to assess writing skills, did not entirely die, however. In fact, there was early research at ETS that would be instrumental in eventually institutionalizing constructed-response formats, although it was a journey of nearly 50 years. The role of ETS in that process has been significant. The chapter on performance assessment by Suzanne Lane and Clement Stone (Lane and Stone 2006) in *Educational Measurement* is an objective measure. Approximately 20% of the chapter's citations were to publications authored by ETS staff. This fact is noteworthy, because the creation of an ETS-like organization had been objected to by Carl Brigham on the grounds that an organization that produced tests would work to preserve the status quo, with little incentive to pursue innovation. As he noted in a letter to Conant (cited in Bennett, Chap. 1, this volume):

> one of my complaints against the proposed organization is that although the word research will be mentioned many times in its charter, the very creation of powerful machinery to do

> more widely those things that are now being done badly will stifle research, discourage new developments, and establish existing methods, and even existing tests, as the correct ones. (p. 6)

His fears were not unreasonable in light of what we know today about the potential for lack of innovation in established organizations (Dougherty and Hardy 1996). However, according to Bennett (2005), from its inception, the ETS Board of Trustees heeded Brigham's concerns, as did the first ETS president (from 1947 until 1970), Henry Chauncey. That climate was favorable to conducting research that would address how to improve and modernize existing tests.[55] Among the many areas of research were investigations related to the scoring of writing. That early research led to a solution to what has been the long-standing problem of operationally scoring essays with acceptable scoring reliability.

Even if the scoring agreement problem was on its way to being solved, it was still the case that tasks requiring a longer constructed response would also take more time and that therefore fewer items could be administered in a given period. With predictive validity as the key metric for evaluating the "validity" of scores, the inclusion of constructed-response tasks continued to encounter resistance. An exception to this trend was the AP program, which relied on constructed-response tasks from its inception. There was also pioneering work on constructed-response assessments early in ETS's history (Frederiksen et al. 1957; Hemphill et al. 1962). However, in both of these cases, the context was very different from the admissions testing case that represented the bulk of ETS business.

Thus a major development toward wider use of constructed-response formats was the evolution of validity theory away from an exclusive focus on predictive considerations. Messick's (1989) work was largely dedicated to expanding the conception of validity to include not only the psychometric attributes of the test, the evidentiary aspect of validation, but also the repercussions that the use of the test could have, the consequential aspect. This broader view did not necessarily endorse the use of one format over another but provided a framework in which constructed-response formats had a greater chance for acceptance.

With the expansion of validity, the doors were opened a bit more, although costs and scalability considerations remained. These considerations were aided by the transition of assessment from paper to computer. The transition to computer-delivered tests at ETS that started in 1985 with the deployment of ACCUPLACER set the stage for the transition of other tests—like the GMAT, GRE, and TOEFL—to digital delivery and the expansion of construct coverage and constructed-response formats, especially for writing and eventually speaking.

[55] A good example can be seen in the evolution of the GRE, a test owned by ETS. In 1992, ETS introduced adaptive testing in the GRE by building on research by Fred Lord and others. In 2011, the GRE was revised again to include, among other changes, a different form of adaptive testing that has proven more robust than the earlier approach. The current adaptive testing approach, a multistage design (Robin et al. 2014), was experimented with much earlier at ETS (Angoff and Huddleston 1958), was extensively researched (Linn et al. 1969), and has since proven to be preferable in an on-demand admissions testing context.

Along the way, there was abundant research and implementation of results in response to the demands resulting from the incorporation of expanded constructs and the use of the computer to support those demands. For example, the psychometric infrastructure for mixed-format designs, including psychometric modeling of polytomous responses, was first used in 1992 (Campbell et al. 1996, p. 113) and developed at ETS. The use of constructed-response formats also required an efficient means of scoring responses captured from booklets. ETS collaborated with subcontractors in developing the necessary technology, as well as developing control procedures to monitor the quality of scoring. Online scoring systems were also developed to accommodate the transition to continuous administration that accompanied computer-based testing. Similarly, automated scoring was first deployed operationally in 1997, when the licensing test for architects developed by ETS became operational (Bejar and Braun 1999; Kenney 1997). The automated scoring of essays was first deployed operationally in 1999 when it was used to score GMAT essays.

Clearly, by the last decade of the twentieth century, the fruits of research at ETS around constructed-response formats were visible. The increasingly ambitious assessments that were being conceived in the 1990s stimulated a rethinking of the assessment design process and led to the conception of ECD (Mislevy et al. 2003). In addition, ETS expanded its research agenda to include the role of assessment in instruction and forms of assessment that, in a sense, are beyond format. Thus the question is no longer one of choice between formats but rather whether an assessment that is grounded in relevant science can be designed, produced, and deployed. That such assessments call for a range of formats and response types is to be expected. The CBAL initiative represents ETS's attempt to conceptualize assessments that can satisfy the different information needs of K–12 audiences with state-of-the-art tasks grounded in the science of student learning, while taking advantage of the latest technological and methodological advances. Such an approach seems necessary to avoid the difficulties that accountability testing has encountered in the recent past.

18.6.1 What Is Next?

If, as Alphonse De Lamartine (1849) said, "history teaches us everything, including the future" (p. 21), what predictions about the future can be made based on the history just presented? Although for expository reasons I have laid the history of constructed-response research at ETS as a series of sequential hurdles that appear to have been solved in an orderly fashion, in reality it is hard to imagine how the story would have unfolded at the time that ETS was founded. While there were always advocates of the use of constructed-response formats, especially in writing, Huddleston's views that writing was essentially verbal ability, and therefore could be measured with multiple-choice verbal items, permeated decision making at ETS.

Given the high stakes associated with admissions testing and the technological limitations of the time, in retrospect, relying on the multiple-choice format arguably

was the right course of action from both the admissions committee's point of view and from the student's point of view. It is well known that James Bryant Conant instituted the use of the SAT at Harvard for scholarship applicants (Lemann 2004), shortly after his appointment as president in 1933, based on a recommendation by his then assistant Henry Chauncey, who subsequently became the first president of ETS.[56] Conant was motivated by a desire to give students from more diverse backgrounds an opportunity to attend Harvard, which in practice meant giving students from other than elite schools a chance to enroll. The SAT, with its curriculum agnostic approach to assessment, was fairer to students attending public high schools than the preparatory-school-oriented essay tests that preceded it. That is, the consequential aspect of validation may have been at play much earlier than Messick's proposal to incorporate consequences of test use as an aspect of validity, and it could be in this sense partially responsible for the long period of limited use into which the constructed-response format fell.

However well-intentioned the use of the multiple-choice format may have been, Frederiksen (1984) claimed that it represented the "real test bias." In doing so, he contributed to fueling the demand for constructed-response forms of assessment. It is possible to imagine that the comforts of what was familiar, the multiple-choice format, could have closed the door to innovation, a fear that had been expressed by Carl Brigham more generally.[57] For companies emerging in the middle of the twentieth century, a far more ominous danger was the potential disruptions that could accrue from a transition to the digital medium that would take place during the second half of the century. The Eastman Kodak Company, known for its film and cameras, is perhaps the best known example of the disruption that the digital medium could bring: It succumbed to the digital competition and filed for bankruptcy in 2012. However, this is not a classic case of being disrupted out of existence,[58] because Kodak invented the first digital camera! The reasons for Kodak's demise are far more nuanced and include the inability of the management team to figure out in time how to operate in a hybrid digital and analog world (Chopra 2013). Presumably a different management team could have successfully transitioned the company to a digital world.[59]

In the testing industry, by contrast, ETS not only successfully navigated the digital transition, *it actually led the transition* to a digital testing environment with the

[56] In 1938, Chauncey and Conant "persuaded all the College Board schools to use the SAT as the main admissions tests for scholarship applicants" (Lemann 2004, p. 8), and in 1942, the written College Board exams were suspended and the SAT became the admissions tests for all students, not just scholarship applicants.

[57] The tension between innovation and "operate and maintain" continues to this day (T. J. Elliot 2014).

[58] Both Foster (1986) and Christensen (1997) discussed the potential of established companies to fall prey to nimble innovating companies.

[59] In the leadership literature (Howell and Avolio 1993), a distinction is made between two styles of leadership: transactional and transformational. Transactional leaders aim to achieve their goals through accountability, whereas transformational leaders aim to achieve their goals by being charismatic and inspirational. It is easy to imagine that different prevailing types of leadership are better or worse at different points in the life of a company.

launching of ACCUPLACER in 1985.[60] The transition to adaptive testing must have been accompanied by a desire to innovate and explore how technology could be used in testing, because in reality, there were probably few compelling business or even psychometric reasons to launch a computer-based placement test in the 1980s. Arguably, the early transition made it possible for ETS to eventually incorporate constructed-response formats into its tests *sooner*, even if the transition was not even remotely motivated by the use of constructed-response formats. By building on a repository of research related to constructed-response formats motivated by validity and fairness considerations, the larger transition to a digital ecosystem did not ultimately prove to be disruptive at ETS and instead made it possible to take advantage of the medium, finally, to deploy assessments containing constructed-response formats and to envision tests as integral to the educational process rather than purely technological add-ons.

In a sense, the response format challenge has been solved: Admissions tests now routinely include constructed-response items, and the assessments developed to measure the Common Core State Standards also include a significant number of constructed-response items. Similarly, NAEP, which has included constructed-response items for some time, is making the transition to digital delivery via tablets. ETS has had a significant role in the long journey. From inception, there was a perspective at ETS that research is critical to an assessment organization (Chauncey, as cited by Bennett, Chap. 1, this volume). Granted that the formative years of the organization were in the hands of enlightened and visionary individuals, it appears that the research that supported the return of constructed-response formats was not prescribed from above but rather the result of *intra*preneurship,[61] or individual researchers largely pursuing their own interests.[62] If this is the formula that worked in the past, it could well continue to work in the future, if we believe De Lamartine. Of course, Santayana argued that history is always written wrong and needs to be rewritten. Complacency about the future, therefore, is not an option—it will still need to be constructed.

Acknowledgments The chapter has greatly benefited from the generous contributions of many colleagues at ETS. Peter Cooper, Robby Kantor, and Mary Fowles shared generously their knowledge of writing assessment. Despina Danos, Alice Sims-Gunzenhauser, and Walt MacDonald did likewise with respect to the history of the Advanced Placement program. Mel Kubota and Marylyn Sudlow shared their knowledge of the history of writing within the SAT admissions program. John Barone, Ina Mullis, and Hillary Persky were my primary sources for NAEP's history. Pan Mollaun and Mary Schedl shared their knowledge on the history of TOEFL. Sydell Carlton generously provided a firsthand account of the milestone study "Factors in Judgments of Writing Ability." And Mike Zieky's long tenure at ETS served to provide me with useful historical background throughout the project. Karen McQuillen and Jason Wagner provided access to the nooks and crannies of the Brigham Library and greatly facilitated tracking down sources. The chronology reported in Table 18.1 would have been nearly impossible to put together without the assistance of Jackie

[60] ETS developed ACCUPLACER for the College Board.

[61] The term that Pinchot (1987) introduced relatively recently to describe within-company entrepreneurship.

[62] N. Elliot (2005, p. 183) wrote instead about "lone wolves."

Briel, Gary Driscoll, Roby Kantor, Fred McHale, and Dawn Piacentino; I'm grateful for their patience. Within the ETS Research division, I benefited from input by Keelan Evanini, Irv Katz, Sooyeon Kim, and Klaus Zechner.

Other individuals were also generous contributors, especially Bert Green, Mary Pommerich, Dan Segal, and Bill Ward, who shared their knowledge of the development of adaptive testing; Nancy Petersen contributed on the origins of automated scoring at ETS. I'm especially grateful to Norbert Elliot for his encouragement. His *On a Scale: A Social History of Writing Assessment in America* was an invaluable source in writing this chapter. He also offered valuable suggestions and sources, many of which I was able to incorporate.

Randy Bennett, Brent Bridgeman, and Don Powers, by virtue of being major contributors to the knowledge base on constructed-response formats at ETS, were ideal reviewers and provided many valuable suggestions in their respective reviews. Doug Baldwin and Larry Stricker offered valuable criticism on an earlier draft. Finally, the manuscript has benefited greatly from the thorough editorial review by Randy Bennett, Jim Carlson, and Kim Fryer. However, it should be clear that I remain responsible for any flaws that may remain.

References

Advanced Placement Program of the College Board. (1980). *An informal history of the AP readings 1956–1976*. New York: The College Board.

Allen, N. L., Holland, P. W., & Thayer, D. T. (2005). Measuring the benefits of examinee-selected questions. *Journal of Educational Measurement, 42*, 27–51. https://doi.org/10.1111/j.0022-0655.2005.00003.x

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*, 341–359. https://doi.org/10.1111/j.1745-3984.2007.00043.x

Anderson, H. A., & Traxler, A. E. (1940). The reliability of the reading of an English essay test: A second study. *The School Review, 48*(7), 521–530.

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test* (Statistical Report No. SR-58-21). Princeton: Educational Testing Service.

Applebee, A. N. (2007). Issues in large-scale writing assessment: Perspectives from the National Assessment of Educational Progress. *Journal of Writing Assessment, 3*(2), 81–98.

Attali, Y. (2010). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement, 38*, 632–644.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e-rater* V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org/

Attali, Y., & Powers, D. E. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended GRE subject test items* (Research Report No. RR-08-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02107.x

Attali, Y., & Powers, D. E. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70*(1), 22–35. https://doi.org/10.1177/0013164409332231

Baker, F. B. (1971). Automation of test scoring, reporting and analysis. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 202–236). Washington, DC: American Council on Education.

Baldwin, D. (2004). A guide to standardized writing assessment. *Educational Leadership, 62*(2), 72–75.

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-responses and other performance assessments*. Princeton: Educational Testing Service.

Bejar, I. I. (1985). *A preliminary study of raters for the Test of Spoken English* (TOEFL Research Report No. 18). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00090.x

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522–532. https://doi.org/10.1037/0021-9010.76.4.522

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–218). Mahwah: Erlbaum.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation: Final report to the National Council of Architectural Registration Boards* (Research Memorandum No. RM-99-02). Princeton: Educational Testing Service.

Bejar, I. I., & Graf, E. A. (2010). Updating the duplex design for test-based accountability in the twenty-first century. *Measurement: Interdisciplinary Research & Perspective, 8*(2), 110–129. https://doi.org/10.1080/15366367.2010.511976

Bejar, I. I., & Whalen, S. J. (2001). *U.S. Patent No. 6,295,439.* Washington, DC: Patent and Trademark Office.

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer. https://doi.org/10.1007/978-1-4613-9690-1

Bejar, I. I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. 19). Princeton: Educational Testing Service.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–82). Mahwah: Erlbaum.

Bennett, R. E. (2005). *What does it mean to be a nonprofit educational measurement organization in the 21st century?* Princeton: Educational Testing Service.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research & Perspective, 8*(2), 70–91. https://doi.org/10.1080/15366367.2010.508686

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–16. https://doi.org/10.1111/j.1745-3992.1998.tb00631.x

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer. https://doi.org/10.1007/978-1-4020-9964-9_3

Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypothesis test. *Journal of Educational Measurement, 32*, 19–36. https://doi.org/10.1111/j.1745-3984.1995.tb00454.x

Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education, 9*, 133–150. https://doi.org/10.1207/s15324818ame0902_3

Bennett, R. E., & Sebrechts, M. M. (1997). A computer-based task for measuring the representational component of quantitative proficiency. *Journal of Educational Measurement, 34*, 64–77. https://doi.org/10.1111/j.1745-3984.1997.tb00507.x

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement*. Hillsdale: Erlbaum.

Bennett, R. E., & Wadkins, J. R. J. (1995). Interactive performance assessment in computer science: The Advanced Placement Computer Science (APCS) Practice System. *Journal of Educational Computing Research, 12*(4), 363–378. https://doi.org/10.2190/5WQA-JB0J-1CM5-4H50

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*, 151–162. https://doi.org/10.1177/014662169001400204

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*, 162–176. https://doi.org/10.1111/j.1745-3984.1997.tb00512.x

Bennett, R. E., Morley, M., Quardt, D., Rock, D. A., Singley, M. K., Katz, I. R., & Nhouyvanishvong, A. (1999). Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning. *Journal of Educational Measurement, 36*, 233–252. https://doi.org/10.1111/j.1745-3984.1999.tb00556.x

Bennett, R. E., Morley, M., & Quardt, D. (2000a). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*, 294–309. https://doi.org/10.1177/01466210022031769

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000b). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education, 13*(3), 303–322. https://doi.org/10.1207/S15324818AME1303_5

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning and Assessment, 8*(8). Retrieved from http://escholarship.bc.edu/jtla/vol8/8

Bernstein, J., de Jong, J., Pisoni, D., & Townshend, D. (2000). Two experiments in automatic scoring of spoken language proficiency. In P. Delcloque (Ed.), *Proceedings of the InSTIL2000: Integrating speech technology in learning* (pp. 57–61). Dundee: University of Abertay.

Bill and Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle: Author.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1988). Comprehensive educational assessment for the states: The Duplex Design. *Educational Evaluation and Policy Analysis, 10*(2), 89–105. https://doi.org/10.1177/014662102237794

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364–375.

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989–2009*. Washington, DC: National Assessment Governing Board.

Braswell, J. S. (1992). Changes in the SAT in 1994. *Mathematics Teacher, 85*(1), 16–21.

Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 167–182). Hillsdale: Erlbaum.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*(1), 1–18. https://doi.org/10.2307/1164948

Braun, H. I., Bejar, I. I., & Williamson, D. M. (2006). Rule-based methods for automated scoring: Applications in a licensing context. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Mahwah: Erlbaum.

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review* (College Board Report No. 83-6). Princeton: Educational Testing Service.

Breland, H. M., & Hart, F. M. (1994). *Defining legal writing: An empirical analysis of the legal memorandum* (Research Report No. 93-06). Newtown: Law School Admission Council.

Breland, H. M., & Jones, R. J. (1988). *Remote scoring of essays* (College Board Report No. 88-03). New York: College Entrance Examination Board.

Breland, H. M., Grandy, J., Rock, D., & Young, J. W. (1984). *Linear models of writing assessments* (Research Memorandum No. RM-84-02). Princeton: Educational Testing Service.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill*. New York: College Entrance Examination Board.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement history examination. *Journal of Educational Measurement, 31*, 275–293. https://doi.org/10.1111/j.1745-3984.1994.tb00447.x

Breland, H. M., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education: Review and framework* (Research Report No. RR-99-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1999.tb01801.x

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer. https://doi.org/10.1007/978-1-4757-3456-0

Bridgeman, B. (1989). *Comparative validity of multiple-choice and free-response items on the Advanced Placement examination in biology* (Research Report No. RR-89-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00327.x

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic outcomes? *Educational Measurement: Issues and Practice, 35*(4), 21–24. https://doi.org/10.1111/emip.12130

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41*, 137–148. https://doi.org/10.1111/j.1745-3984.2004.tb01111.x

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37–50. https://doi.org/10.1111/j.1745-3984.1994.tb00433.x

Bridgeman, B., & McHale, F. J. (1996). *Gender and ethnic group differences on the GMAT Analytical Writing Assessment* (Research Report No. RR-96-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1996.tb01680.x

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*, 313–329.

Bridgeman, B., Morgan, R., & Wang, M.-M. (1996a). *Reliability of Advanced Placement examinations* (Research Report No. RR-96-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1996.tb01681.x

Bridgeman, B., Morgan, R., & Wang, M.-M. (1996b). *The reliability of document-based essay questions on Advanced Placement history examinations* (Research Report No. RR-96-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1996.tb01683.x

Bridgeman, B., Morgan, R., & Wang, M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement, 34*, 273–286. https://doi.org/10.1111/j.1745-3984.1997.tb00519.x

Bridgeman, B., Trapani, C., & Bivens-Tatum, J. (2011). Comparability of essay question variants. *Assessing Writing, 16*(4), 237–255. https://doi.org/10.1016/j.asw.2011.06.002

Bridgeman, B., Powers, D. E., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91–108. https://doi.org/10.1177/0265532211411078

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44*, 131–155. https://doi.org/10.1111/j.1745-3984.2007.00031.x

Brunswik, E. (1952). The conceptual framework of psychology. In *International encyclopedia of unified science* (Vol. 1, No. 10). Chicago: University of Chicago Press.

Burke, V. M. (1961). A candid opinion on lay readers. *The English Journal, 50*(4), 258–264. https://doi.org/10.2307/810913

Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1997). *Automatic scoring of Advanced Placement Biology essays* (Research Report No. RR-97-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1997.tb01743.x

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT Analytical Writing Assessment essays* (Research Report No. RR-98-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01764.x

Burstein, J., Wolff, S., & Lu, C. (1999). Using lexical semantic techniques to classify free responses. In N. Ide & J. Veronis (Eds.), *The depth and breadth of semantic lexicons* (pp. 227–244). Dordrecht: Kluwer Academic. https://doi.org/10.1007/978-94-017-0952-1_11

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation. *AI Magazine, 25*(3), 27–36.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current application and new directions* (pp. 55–67). New York: Routledge.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (Research Memorandum No. RM-00-06). Princeton: Educational Testing Service.

Cameron, C. A., Beemsterboer, P. L., Johnson, L. A., Mislevy, R. J., Steinberg, L. S., & Breyer, F. J. (2000). A cognitive task analysis for dental hygiene. *Journal of Dental Education, 64*(5), 333–351.

Camp, R. (1985). The writing folder in post-secondary assessment. In P. J. A. Evans (Ed.), *Directions and misdirections in English evaluation* (pp. 91–99). Ottawa: Canadian Council of Teachers of English.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 183–212). Hillsdale: Erlbaum.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. https://doi.org/10.1037/h0046016

Campbell, J. R., Donahue, P. L., Reese, C. M., & Phillips, G. W. (1996). *NAEP 1994 Reading report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. https://doi.org/10.1093/applin/1.1.1

Chalifour, C., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26*, 120–132. https://doi.org/10.1111/j.1745-3984.1989.tb00323.x

Chopra, A. (2013). *The dark side of innovation. St*. Johnsbury: Brigantine.

Christensen, C. M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail*. Boston: Harvard Business School Press.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage. https://doi.org/10.4135/9781412985918

Clark, J. L. D., & Swinton, S. S. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report No. 07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1980.tb01230.x

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*, 310–324. https://doi.org/10.1177/01466210022031778

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413–432.

Cleary, A. T. (1966). *Test bias: Validity of the Scholastic Aptitude Test for Negro and White students in integrated colleges* (Research Bulletin No. RB-66-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1966.tb00529.x

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based simulations. In E. L. Mancall & E. G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 139–149). Evanston: American Board of Medical Specialties.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 271–302). Washington, DC: American Council on Education. https://doi.org/10.3102/00028312005001099

Coffman, W. E., & Kurfman, D. G. (1966). *Single score versus multiple score reading of the American History Advanced Placement examination (Research Bulletin No. RB-66-22)*. Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1966.tb00359.x

Coffman, W. E., & Kurfman, D. G. (1968). A comparison of two methods of reading essay examinations. *American Educational Research Journal, 5*(1), 99–107.

College Entrance Examination Board. (2011). *AP World History: Course and exam description*. New York: Author.

College Entrance Examination Board. (1942). *Forty-second annual report of the Executive Secretary*. New York: Author.

Committee on the Objectives of a General Education in a Free Society. (1945). *General education in a free society: Report of the Harvard committee*. Cambridge, MA: Harvard University Press.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing* (pp. 3–33). Urbana: National Council of Teachers of English.

Coward, A. F. (1950). *The method of reading the Foreign Service Examination in English composition* (Research Bulletin No. RB-50-57). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00680.x

Coward, A. F. (1952). The comparison of two methods of grading English compositions. *Journal of Educational Research, 46*(2), 81–93. https://doi.org/10.1080/00220671.1952.10882003

Cowden, D. J. (1946). An application of sequential sampling to testing students. *Journal of the American Statistical Association, 41*(236), 547–556. https://doi.org/10.1080/01621459.1946.10501897

Crisp, V. (2010). Towards a model of the judgment processes involved in examination marking. *Oxford Review of Education, 36*, 1–21. https://doi.org/10.1080/03054980903454181

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement over a decade (Proceedings of the 1979 ETS Invitational Conference)* (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. https://doi.org/10.1037/h0040957

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). Generalizability analysis for educational assessments. *Educational and Psychological Measurement, 57*(3), 373–399. https://doi.org/10.1177/0013164497057003001

Croon Davis, C., Stalnaker, J. M., & Zahner, L. C. (1959). Review of the Sequential Tests of Educational Progress: Writing. In O. K. Buros (Ed.), *The fifth mental measurement yearbook*. Lincoln: Buros Institute of Mental Measurement.

Cumming, A., Kantor, R., Powers, D. E., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series No. 18). Princeton: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

D'Angelo, F. (1984). Nineteenth-century forms/modes of discourse: A critical inquiry. *College Composition and Communication, 35*(1), 31–42. https://doi.org/10.2307/357678

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106. https://doi.org/10.1037/h0037613

Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 313–371). Mahwah: Erlbaum.

Deane, P. (2012). Rethinking K–12 writing assessment. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 87–100). New York: Hampton Press.

Deane, P. (2013a). Covering the construct: An approach to automated essay scoring motivated by a socio-cognitive framework for defining literacy skills. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current application and new directions* (pp. 298–312). New York: Routledge.

Deane, P. (2013b). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research, 2*(2), 151–177. https://doi.org/10.17239/jowr-2010.02.02.4

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (Research Report No. RR-08-55). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02141.x

Deane, P., Sabatini, J., & Fowles, M. (2012). Rethinking k-12 writing assessment to support best instructional practices. In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 81–102). Anderson: The WAC Clearinghouse and Parlor Press.

De Lamartine, A. (1849). *History of the French Revolution of 1848*. London: Bohn.

Diederich, P. B. (1957). *The improvement of essay examinations* (Research Memorandum No. RM-57-03). Princeton: Educational Testing Service.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin No. RB-61-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1961.tb00285.x

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 135–166). Hillsdale: Erlbaum.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program examinations. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement program examinations* (Research Report No. RR-03-27; pp. 79–118). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01919.x

Dorans, N. J., Holland, P., & Pommerich, M. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer. https://doi.org/10.1007/978-0-387-49771-6

Dougherty, D., & Hardy, C. (1996). Sustained product innovation in large, mature organizations: Overcoming innovation-to-organization problems. *Academy of Management Journal, 39*(5), 1120–1153. https://doi.org/10.2307/256994

Downey, M. T. (1965). *Ben D. Wood, educational reformer*. Princeton: Educational Testing Service.

Driscoll, G., Hatfield, L. A., Johnson, A. A., Kahn, H. D., Kessler, T. E., Kuntz, D., et al. (1999). *U.S. Patent No. 5,987,302*. Washington, DC: U.S. Patent and Trademark Office.

Du Bois, P. H. (1970). *A history of psychological testing*. Boston: Allyn and Bacon.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1987). TOEFL from a communicative viewpoint on language proficiency: A working paper. In R. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 1–150). Norwood: Ablex.

Dwyer, C. A. (1994). *Development of the knowledge base for the PRAXIS III: Classroom performance assessments assessment criteria*. Princeton: Educational Testing Service.

Dyer, H. S. (1954, January 5). *A common philosophy for the Test of Developed Ability*. Princeton: Educational Testing Service.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407–424. https://doi.org/10.1007/BF02288803

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement, 22*(1), 15–25.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*(4), 460–475.

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on e-mail. *Assessing Writing, 1*(1), 91–107. https://doi.org/10.1016/1075-2935(94)90006-X

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Lang.

Elliot, N. (2014). *Henry Chauncey: An American life*. New York: Lang.

Elliot, T. J. (2014). Escaping gravity: Three kinds of knowledge as fuel for innovation in an operate and maintain company. In K. Pugh (Ed.), *Smarter innovation: Using interactive processes to drive better business results* (pp. 3–12). Peoria: Ark Group.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Enright, M. K., & Gitomer, D. (1989). *Toward a description of successful graduate students* (Research Report No. RR-89-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00335.x

Enright, M. K., & Powers, D. E. (1991). *Validating the GRE Analytical Ability Measure against faculty ratings of analytical reasoning skills* (Research Report No. RR-90-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1990.tb01358.x

Enright, M. K., Rock, D. A., & Bennett, R. E. (1998). Improving measurement for graduate admissions. *Journal of Educational Measurement, 35*, 250–267. https://doi.org/10.1111/j.1745-3984.1998.tb00538.x

Enright, M. K., Grabe, W., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series No. 17). Princeton: Educational Testing Service.

Evanini, K., Xie, S., & Zechner, K. (2013). *Prompt-based content scoring for automated spoken language assessment.* Paper presented at the Eighth Workshop on the Innovative Use of NLP for Building Educational Applications, Atlanta.

Fife, J. H. (2013). *Automated scoring of mathematics tasks in the Common Core Era: Enhancements to m-rater™ in support of CBAL mathematics and the Common Core Assessments* (Research Report No. RR-13-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02333.x

Finlayson, D. S. (1951). The reliability of the marking of essays. *British Journal of Educational Psychology, 21*, 126–134. https://doi.org/10.1111/j.2044-8279.1951.tb02776.x

Finley, C., & Berdie, F. S. (1970). *The National Assessment approach to exercise development*. Ann Arbor, MI: National Assessment of Educational Progress. Retrieved from ERIC database. (ED 067402)

Foster, R. N. (1986). *Innovation: The attacker's advantage*. New York: Summit Books. https://doi.org/10.1007/978-3-322-83742-4

Fowles, M. (2012). Writing assessments for admission to graduate and professional programs: Lessons learned and a note for the future. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 135–148). New York: Hampton Press.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202. https://doi.org/10.1037/0003-066X.39.3.193

Frederiksen, N., Saunders, D. R., & Wand, B. (1957). The in-basket test. *Psychological Monographs, 71*(438). https://doi.org/10.1037/h0093706

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York: Longman.

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8*(1), 27–41. https://doi.org/10.1207/S15326977EA0801_02

General Education in School and College. (1952). *A committee report by members of the faculties of Andover, Exeter, Lawrenceville, Harvard, Princeton, and Yale*. Cambridge: Harvard University Press.

Gentile, C. A. (1992). *NAEP's 1990 pilot portfolio study: Exploring methods for collecting students' school-based writing*. Washington, DC: National Center for Education Statistics.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gitomer, D. H. (2007). *The impact of the National Board for Professional Teaching Standards: A review of the research* (Research Report No. RR-07-33). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02075.x

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521. https://doi.org/10.1037/h0049294

Glaser, R. (1994). Criterion-referenced tests: Part I. Origins. *Educational Measurement: Issues and Practice, 13*(4), 9–11. https://doi.org/10.1111/j.1745-3992.1994.tb00562.x

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Golub-Smith, M., Reese, C., & Steinhaus, K. (1993). *Topic and topic type comparability on the Test of Written English* (TOEFL Research Report No. 42). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01521.x

Graf, E. A. (2009). *Defining mathematics competency in the service of cognitively based assessment for Grades 6 through 8* (Research Report No. RR-09-42). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02199.x

Graf, E. A., & Fife, J. (2012). Difficulty modeling and automatic item generation of quantitative items: Recent advances and possible next steps. In M. Gierl & S. Haladyna (Eds.), *Automated item generation: Theory and practice* (pp. 157–180). New York: Routledge.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley. https://doi.org/10.1037/13240-000

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology-Research and Practice, 11*(3), 385–398. https://doi.org/10.1037/0735-7028.11.3.385

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport: American Council on Education and Praeger.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion referenced tests. *Journal of Educational Measurement, 10*, 159–170. https://doi.org/10.1111/j.1745-3984.1973.tb00793.x

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport: American Council on Education and Praeger.

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*(6), 438–456. https://doi.org/10.1037/h0040736

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000—Writing: Composition, community, and assessment* (TOEFL Monograph Series No. MS-05). Princeton: Educational Testing Service.

Hansen, E. G., & Mislevy, R. J. (2008). *Design patterns for improving accessibility for test takers with disabilities* (Research Report No. RR-08-49). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02135.x

Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*(1), 5–40. https://doi.org/10.1207/s15326977ea0501_1

Hemphill, J. K., Griffiths, D. E., & Frederiksen, N. (1962). *Administrative performance and personality: A study of the principal in a simulated elementary school*. New York: Teachers College.

Hick, W. E. (1951). Information theory and intelligence tests. *British Journal of Psychology 4(3)*, 157–164.

Hines, S. (2010). *Evidence-centered design: The TOEIC speaking and writing tests* (TOEIC Compendium Study No. 7). Princeton: Educational Testing Service.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2), 1–49.

Howell, J., & Avolio, B. J. (1993). Transformational leadership, transactional leadership, locus of control, and support for innovation: Key predictors of consolidated-business-unit performance. *Journal of Applied Psychology, 78*(6), 891–902. https://doi.org/10.1037/0021-9010.78.6.891

Hubin, D. R. (1988). *The Scholastic Aptitude Test: Its development and introduction, 1900–1947*. Retrieved from http://darkwing.uoregon.edu/~hubin/

Huddleston, E. M. (1954). The measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques. *Journal of Experimental Education, 22*(3), 165–207. https://doi.org/10.1080/00220973.1954.11010477

Huff, K., & Plake, B. S. (2010). Evidence-centered assessment design in practice. *Applied Measurement in Education, 23*(4), 307–309. https://doi.org/10.1080/08957347.2010.510955

Huff, K., Alves, C., Pellegrino, J., & Kaliski, P. (2012). Using evidence-centered design task models in automatic item generation. In M. Gierl & S. Haladyna (Eds.), *Automated item generation: Theory and practice* (pp. 102–118). New York: Routledge.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201–213. https://doi.org/10.2307/358160

Huot, B., & Neal, M. (2006). Writing assessment: A techno-history. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 417–432). New York: Guilford Press.

Hutt, M. L. (1947). A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. *Journal of Consulting Psychology, 11*, 93–103. https://doi.org/10.1037/h0056579

Joe, J. N., Park, Y. S., Brantley, W., Lapp, M., & Leusner, D. (2012, April). *Examining the effect of prompt complexity on rater behavior: A mixed-methods study of GRE Analytical Writing Measure Argument prompts*. Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, British Columbia, Canada.

Jones, V., & Brown, R. H. (1935). Educational tests. *Psychological Bulletin, 32*(7), 473–499. https://doi.org/10.1037/h0057574

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160. https://doi.org/10.1177/014662168200600201

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger.

Kaplan, R. (1992). *Using a trainable pattern-directed computer program to score natural language item responses* (Research Report No. RR-91-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01398.x.

Kaplan, R. M., & Bennett, R. E. (1994). *Using the free-response scoring tool to automatically score the formulating-hypotheses item* (Research Report No. RR-94-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1994.tb01581.x

Kaplan, R. M., Burstein, J., Trenholm, H., Lu, C., Rock, D., Kaplan, B., & Wolff, C. (1995). *Evaluating a prototype essay scoring procedure using off-the shelf software* (Research Report No. RR-95-21). Princeton: Educational Testing Service. https://dx.doi.org/10.1002/j.2333-8504.1995.tb01656.x

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics, 23*(3), 254–278. https://doi.org/10.3102/10769986023003254

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*, 39–57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Kenney, J. F. (1997). New testing methodologies for the Architect Registration Examination. *CLEAR Exam Review, 8*(2), 23–28.

Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*, 53–76. https://doi.org/10.1111/j.1745-3984.2006.00004.x

Kim, S., & Walker, M. (2012). Determining the anchor composition for a mixed-format test: Evaluation of subpopulation invariance of linking functions. *Applied Measurement in Education, 25*(2), 178–195. https://doi.org/10.1080/08957347.2010.524720

Kim, S., Walker, M. E., & McHale, F. J. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*, 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x

Kirsch, I., Braun, H. I., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future*. Princeton: Educational Testing Service.

Koenig, J. A. (2011). *Assessing 21st century skills: Summary of a workshop*. Washington, DC: National Academies Press.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport: American Council on Education and Praeger.

Kuntz, D., Cody, P., Ivanov, G., & Perlow, J. E. (2006). *U.S. Patent No. 8,374,540 B2*. Washington, DC: Patent and Trademark Office.

Lacy, T. (2010). Examining AP: Access, rigor, and revenue in the history of the Advanced Placement program. In D. R. Sadler, G. Sonnert, R. H. Tai, & K. Klopfenstein (Eds.), *AP: A critical examination of the Advanced Placement program* (pp. 17–48). Cambridge, MA: Harvard Education Press.

Lane, E. S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport: American Council on Education and Praeger.

Lawrence, I., Rigol, G., Van Essen, T., & Jackson, C. A. (2002). *A historical perspective on the SAT 1926–2001* (Research Report No. 2002-7). New York: College Entrance Examination Board.

Leacock, C., & Chodorow, M. (2003). c-rater: Scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405. https://doi.org/10.1023/A:1025779619903

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus, and Giroux.

Lemann, N. (2004). A history of admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standarized testing in univesity admissions* (pp. 5–14). New York: RoutledgeFalmer.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386. https://doi.org/10.1177/014662169001400404

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97*(2), 460–468.

Linacre, J. M. (2010). Facets Rasch measurement, version 3.67.1 [Computer software]. Chicago: Winsteps.com.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

Linn, R. L. (1996). Work readiness assessment: Questions of validity. In B. L. Resnick & J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment* (pp. 245–266). San Francisco: Jossey-Bass.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5–8. https://doi.org/10.1111/j.1745-3992.1994.tb00778.x

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River: Prentice Hall.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29*(1), 29–146. https://doi.org/10.1177/001316446902900109

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21. https://doi.org/10.3102/0013189X020008015

Liu, L., Rogat, A., & Bertling, M. (2013). *A CBAL science model of cognition: Developing a competency model and learning progressions to support assessment development* (Research Report No. RR-13-29). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02336.x

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*(2), 19–28. https://doi.org/10.1111/emip.12028

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*, 215–233. https://doi.org/10.1002/tea.21299

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–66). Urban: National Council of Teachers of English.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 653–694. https://doi.org/10.2466/pr0.1957.3.3.635

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72*, 143–155. https://doi.org/10.1037/h0021704

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19*(3), 171–200. https://doi.org/10.2307/1165293

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph, 17*(7).

Lord, F. M. (1965). *Item sampling in test theory and in research design* (Research Bulletin No. RB-65-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1965.tb00968.x

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

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234–250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response assessments* (Unpublished doctoral dissertation). Pittsburgh: Carnegie Mellon University.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature*. (Report No. 88-8). New York: College Entrance Examination Board.

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 661–699). Westport: American Council on Education and Praeger.

McClellan, C. A. (2010). *Constructed-response scoring—Doing it right* (R&D Connections No. 13). Princeton: Educational Testing Service.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of evidence*. Minneapolis: University of Minnesota Press. https://doi.org/10.1037/11281-000

Messick, S. (1972). Beyond structure: In search of functional models of psychological process. *Psychometrika, 37*, 357–375. https://doi.org/10.1007/BF02291215

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillian.

Messick, S. (1990). *Validity of test interpretation and use* (Research Report No. RR-90-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1990.tb01343.x

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61–74). Hillsdale: Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Messick, S., & Kogan, N. (1966). Personality consistencies in judgment: Dimensions of role constructs. *Multivariate Behavioral Research, 1*, 165–175. https://doi.org/10.1207/s15327906mbr0102_3

Messick, S., Beaton, A. E., Lord, F. M., Baratz, J. C., Bennett, R. E., Duran, R. P., et al. (1983). *National assessment of educational progress reconsidered: A new design for a new era* (NAEP Report No. 83-1). Princeton: Educational Testing Service.

Miller, G. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences, 7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). Amsterdam, the Netherlands: Kluwer. https://doi.org/10.1007/0-306-47531-6_4

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine, 178*(Suppl. 1), 107–114. https://doi.org/10.7205/MILMED-D-13-00213

Mislevy, R. J., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). London: Routledge.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15*(3–4), 335–374. https://doi.org/10.1016/S0747-5632(99)00027-8

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002a). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Mahwah: Erlbaum.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002b). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363–390. https://doi.org/10.1016/S0747-5632(99)00027-8

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J., Steinberg, L., Almond, R. G., & Lucas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–82). Mahwah: Erlbaum.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., et al. (2014). *Psychometrics considerations in game-based assessment*. Retrieved from http://www.instituteofplay.org/wp-content/uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf

Mislevy, R. J., Corrigan, S., Oranje, A., Dicerbo, K., Bauer, M. I., von Davier, A. A., & Michael, J. (2016). Psychometrics for game-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 23–48). Washington, DC: National Council on Measurement in Education.

Moses, T. P., Yang, W.-L., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement, 44*, 157–178. https://doi.org/10.1111/j.1745-3984.2007.00032.x

Mullis, I. V. S. (1980). *Using the primary trait system for evaluating writing*. Washington, DC: National Assessment of Educational Progress.

Mullis, I. V. S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement: Issues and Practice, 3*(1), 16–18. https://doi.org/10.1111/j.1745-3992.1984.tb00728.x

Myers, A. E., McConville, C. B., & Coffman, W. E. (1966). Simplex structure in the grading of essay tests. *Educational and Psychological Measurement, 26*(1), 41–54.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (CSE Technical Report No. 402). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1995). *Reader calibration and its potential role in equating for the Test of Written English* (Research Report No. RR-95-40). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01674.x

Nagy, W. E., & Scott, J. (2000). Vocabulary processes. In M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah: Erlbaum.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform, a report to the Nation and the Secretary of Education, United States Department of Education*. Retrieved from the Department of Education website: http://www2.ed.gov/pubs/NatAtRisk/index.html

National Research Council. (2010). *Best practices for state assessment systems. Part 1, Summary of a workshop*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12906

Norton, L. S. (1990). Essay-writing: What really counts? *Higher Education, 20*, 411–442. https://doi.org/10.1007/BF00136221

Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). An adaptive approach to group-score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 371–390). New York: CRC Press.

Owens, K. M. (2006). *Use of the GMAT Analytical Writing Assessment: Past and present* (Research Report No. 07-01). McLean: GMAC.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561–565.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics in Medicine, 27*, 341–384. https://doi.org/10.3102/10769986027004341

Paul, S. R. (1981). Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology, 34*, 213–223. https://doi.org/10.1111/j.2044-8317.1981.tb00630.x

Pearlman, M. (2008a). Finalizing the test blueprint. In C. A. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). New York: Routledge.

Pearlman, M. (2008b). The design architecture of NBPTS certification assessments. In L. Ingvarson & J. Hattie (Eds.), *Advances in program evaluation: Vol 11. Assessing teachers for professional certification: The first decade of the National Board of Professional Teaching Standards* (pp. 55–91). Bingley: Emerald Group. https://doi.org/10.1016/S1474-7863(07)11003-6

Pellegrino, J. W., & Glaser, R. (1980). Components of inductive reasoning. In P. A. Federico, R. E. Snow, & W. E. Montague (Eds.), *Aptitude, learning, and instruction: Cognitive process analyses* (pp. 177–217). Hillsdale: Erlbaum.

Persky, H. (2012). Writing assessment in the context of the National Assessment of Educational Progress. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 69–86). New York: Hampton Press.

Pinchot, G. (1987). Innovation through intrapreneuring. *Research Management, 30*(2), 14–19.

Popper, K. R. (1992). *The logic of scientific discovery*. London: Routledge. (Original work published 1959).

Powers, D. E. (1984). *Considerations for developing measures of speaking and listening* (Research Report No. RR-84-18). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1984.tb00058.x

Powers, D. E. (2010). *The case for a comprehensive, four-skills assessment of English-language proficiency* (R&D Connections No. 14). Princeton: Educational Testing Service.

Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education, 12*(3), 257–279. https://doi.org/10.1207/S15324818AME1203_3

Powers, D. E., & Dwyer, C. A. (2003). *Toward specifying a construct of reasoning* (Research Memorandum No. RM-03-01). Princeton: Educational Testing Service.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. *Journal of Higher Education, 58*(6), 658–682. https://doi.org/10.2307/1981103

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects of essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220–233. https://doi.org/10.1111/j.1745-3984.1994.tb00444.x

Powers, D. E., Kubota, M. Y., Bentley, J., Farnum, M., Swartz, R., & Willard, A. E. (1998). *Qualifying readers for the Online Scoring Network: Scoring argument essays* (Research Report No. RR-98-28). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01777.x

Powers, D. E., Fowles, M. E., & Welsh, C. (1999a). *Further validation of a writing assessment for graduate admissions* (Research Report No. RR-99-18). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1999.tb01816.x

Powers, D. E., Schedl, M. A., Leung, S. W., & Butler, F. (1999b). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*(4), 399–425.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407–425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W. H. (2009). *The TOEIC speaking and writing tests: Relations to test-taker perceptions of proficiency in English* (TOEIC Research Report No. 4). Princeton: Educational Testing Service.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment, 6*(3). Retrieved from http://www.jtla.org/

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237–265). Cresskill: Hampton Press.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of e-rater* (Research Report No. RR-09–01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02158.x

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30*(1), 39–56. https://doi.org/10.1007/BF02289746

Reid-Green, K. S. (1990). A high speed image processing system. *IMC Journal, 26*(2), 12–14.

Riconscente, M. M., Mislevy, R. J., & Hamel, L. (2005). *An introduction to PADI task templates* (Technical Report No. 3). Menlo Park: SRI International.

Robin, F., Steffen, M., & Liang, L. (2014). The implementation of the GRE Revised General Test as a multistage test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 325–341). Boca Raton: CRC Press.

Rupp, A. A., Levy, R., Dicerbo, K. E., Crawford, A. V., Calico, T., Benson, M., et al. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4*, 1–102.

Russell, M. (2006). *Technology and assessment: The tale of two interpretations*. Greenwich: Information Age.

Sachse, P. P. (1984). Writing assessment in Texas: Practices and problems. *Educational Measurement: Issues and Practice, 3*(1), 21–23. https://doi.org/10.1111/j.1745-3992.1984.tb00731.x

Sakaguchi, K., Heilman, M., & Madnani, N. (2015). Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp. 1049–1054). Retrieved from https://aclweb.org/anthology/N/N15/N15-1000.pdf

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34* (4, Whole Pt. 2). https://doi.org/10.1007/BF03372160

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment project* (NCES 2005-457). Washington, DC: U.S. Government Printing Office.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport: American Council on Education and Praeger.

Schmitt, A., Holland, P., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale: Erlbaum.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Agreement between expert-system and human raters on complex constructed-response quantitative items. *Journal of Applied Psychology, 76*, 856–862. https://doi.org/10.1037/0021-9010.76.6.856

Sebrechts, M. M., Enright, M. K., Bennett, R. E., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction, 14*(3), 285–343. https://doi.org/10.1207/s1532690xci1403_2

Sheehan, K. M., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16*, 65–76. https://doi.org/10.1177/014662169201600108

Sheehan, K. M., & O'Reilly, T. (2011). *The CBAL reading assessment: An approach for balancing measurement and learning goals* (Research Report No. RR-11-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02257.x

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2–16. https://doi.org/10.3102/0013189X020007002

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters background and training on the reliability of direct writing tests. *Modern Language Journal, 76*(1), 27–33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced scoring guide for selected short-text mathematics items* (Field Test 2014). Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/10/Smarter-Balanced-Scoring-Guide-for-Selected-Short-Text-Mathematics-Items.pdf

Song, Y., Heilman, M., Klebanov Beigman, B., & Deane, P. (2014, June). *Applying argumentation schemes for essay scoring*. Paper presented at the First Workshop on Argumentation Mining, Baltimore, Maryland. https://doi.org/10.3115/v1/W14-2110

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). *Assessing written communication in higher education: Review and recommendations for next-generation assessment* (Research Report No. RR-14-37). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12035

Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition*. New York: Macmillan.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Stansfield, C. (1986a). A history of the Test of Written English: The developmental year. *Language Testing, 3*, 224–334. https://doi.org/10.1177/026553228600300209

Stansfield, C. (1986b). *Toward communicative competence testing: Proceedings of the second TOEFL Invitational Conference* (TOEFL Research Report No. 21). Princeton: Educational Testing Service.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84*(4), 353–378. https://doi.org/10.1037/0033-295X.84.4.353

Stetcher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford: Stanford University, Stanford Center for Opportunity Policy in Education.

Stocking, M. (1969). *Short tailored tests* (Research Bulletin No. RB-69-63). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1969.tb00741.x

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292. https://doi.org/10.1177/014662169301700308

Stricker, L. J. (1982). Interpersonal competence instrument: Development and preliminary findings. *Applied Psychological Measurement, 6*, 69–81. https://doi.org/10.1177/014662168200600108

Stricker, L. J., & Rock, D. A. (1990). Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences, 11*(8), 833–839. https://doi.org/10.1016/0191-8869(90)90193-U

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*(4), 665–693. https://doi.org/10.1111/j.1559-1816.2004.tb02564.x

Sukkarieh, J. Z., & Bolge, E. (2008). Leveraging c-rater's automated scoring capability for providing instructional feedback for short constructed responses. In P. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent tutoring systems: Vol. 5091. Proceedings of the 9th international conference on intelligent tutoring systems, 2008, Montreal, Canada* (pp. 779–783). New York: Springer. https://doi.org/10.1007/978-3-540-69132-7_106

Suto, I., & Nadas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education, 24*(3), 335–377. https://doi.org/10.1080/02671520801945925

Suto, I., Nadas, R., & Bell, J. (2009). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*, 21–51. https://doi.org/10.1080/02671520902721837

Taylor, C. A., & Angelis, P. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. Enright, & J. Jamieson (Eds.), *Bulding a validity argument for the Test of English as a Foreign Language* (pp. 27–54). New York: Routledge. https://doi.org/10.1007/BF02288939

Taylor, C. W. (1947). A factorial study of fluency in writing. *Psychometrika, 12*, 239–262. https://doi.org/10.1007/BF02288939

Thomas, W. H., Storms, B. A., Sheingold, K., Heller, J. I., Paulukonis, S. T., & Nunez, A. M. (1998). *California Learning Assessment System: Portfolio assessment research and development project: Final report*. Princeton: Educational Testing Service.

Thurstone, L. L. (1926). *The nature of intelligence*. New York: Harcourt, Brace.

TOEFL. (2011). TOEFL program history. *TOEFL iBT Research Insight, 1*(6).

Torgerson, W. S., & Green, B. F. (1952). A factor analysis of English essay readers. *Journal of Educational Psychology, 43*(6), 354–363. https://doi.org/10.1037/h0052471

Toulmin, S. E. (1958). *The uses of argument*. New York: Cambridge University Press.

Traxler, A. E. (1951). Administering and scoring the objective test. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 329–416). Washington, DC: American Council on Education.

Traxler, A. E. (1954). Impact of machine and devices on developments in testing and related fields. In *Proceedings of the 1953 invitational conference on testing problems* (pp. 139–146). Princeton: Educational Testing Service.

Tryon, R. C. (1935). A theory of psychological components—An alternative to "mathematical factors." *Psychological Review, 42*, 425–454. https://doi.org/10.1037/h0058874

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review, 71*(6), 528–530. https://doi.org/10.1037/h0047061

Turnbull, W. W. (1949). Influence of cultural background on predictive test scores. In *Proceedings of the ETS invitational conference on testing problems* (pp. 29–34). Princeton: Educational Testing Service.

Turnbull, W. W. (1968). Relevance in testing. *Science, 160*, 1424–1429. https://doi.org/10.1126/science.160.3835.1424

U.S. Congress & Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (No. OTA-SET-519). Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor, Secretary's Commission on Achieving Necessary Skills. (1991). *What work requires of schools: A SCANS report for America 2000*. Washington, DC: Author.

Valentine, J. A. (1987). *The College Board and the school curriculum: A history of the College Board's influence on the substance and standards of American education, 1900–1980*. New York: College Entrance Examination Board.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood: Albex.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*(1), 49–71. https://doi.org/10.1111/bmsp.12003

Wainer, H., & Robinson, D. H. (2007). Fumiko Samejima. *Journal of Educational and Behavioral Statistics, 32*(2), 206–222. https://doi.org/10.3102/1076998607301991

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103–118. https://doi.org/10.1207/s15324818ame0602_1

Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of NAACL-HLT 2013* (pp. 814–819). Atlanta: Association of Computational Lingustics.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6*, 1–11. https://doi.org/10.1177/014662168200600101

Ward, W. C. (1984). Using microcomputers to administer tests. *Educational Measurement: Issues and Practice, 3*(2), 16–20. https://doi.org/10.1111/j.1745-3992.1984.tb00744.x

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 271–282.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine scorable forms of a test of scientific thinking. *Journal of Educational Measurement, 17*, 11–29. https://doi.org/10.1111/j.1745-3984.1980.tb00811.x

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report No. 73-1). Minneapolis: University of Minnesota.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400–409. https://doi.org/10.2307/357792

White House. (2009, March 10). *Remarks by the president to the Hispanic Chamber of Commerce on a complete and competitive American education*. Retrieved from https://www.whitehouse.gov/the_press_office/Remarks-of-the-President-to-the-United-States-Hispanic-Chamber-of-Commerce/

Whitely, S. E., & Dawis, R. V. (1974). Effects of cognitive intervention on latent ability measured from analogy items. *Journal of Educational Psychology, 66*, 710–717.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703–713.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah: Erlbaum.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. W., Bolton, S., Feltovich, B., & Welch, C. (1993). *A comparison of word-processed and handwritten essays from a standardized writing assessment* (Research Report No. 93-8). Iowa City: American College Testing.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465–492. https://doi.org/10.1177/0741088398015004002

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *Journal of Technology, Learning, and Assessment, 10*(1.) Retrieved from http://www.jtla.org/

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology, 92*(2), 577–585. https://doi.org/10.1037/0021-9010.92.2.577

Wood, R. (1973). Response-contingent testing. *Review of Educational Research, 43*(4), 529–544. https://doi.org/10.3102/00346543043004529

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT* (pp. 180–189). Stroudsburg: Association for Computational Linguistics.

Zechner, K., Bejar, I. I., & Hemat, R. (2007a). *Toward an understanding of the role of speech recognition in nonnative speech assessment* (TOEFL iBT Research Series No. 2). Princeton: Educational Testing Service.

Zechner, K., Higgins, D., & Xi, X. (2007b, October). *SpeechRater: A construct-driven approach to scoring spontaneous non-native speech*. Paper presented at the 2007 Workshop of the International Speech Communication Association (ISCA) Special Interest Group on Speech and Language Technology in Education (SLaTE), Farmington.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009a). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*, 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Zechner, K., Sabatini, J., & Chen, L. (2009b). Automatic scoring of children's read-aloud text passages and word lists. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 10–18). Stroudsburg: Association for Computational Linguistics.

Zechner, K., Evanini, K., & Laitusis, C. (2012, September). *Using automatic speech recognition to assess the reading proficiency of a diverse sample of middle school students.* Paper presented at the Interspeech Workshop on Child, Computer Interaction, Portland.

Zhang, L. Y., Powers, D. E., Wright, W., & Morgan, R. (2003). *Applying the Online Scoring Network (OSN) to Advanced Placement Program (AP) tests* (Research Report No. RR-03-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01904.x

Zhang, T., Mislevy, R. J., Haertel, G., Javitz, H., & Hansen, E. G. (2010). *A Design pattern for a spelling assessment for students with disabilities* (Technical Report No. 2). Menlo Park: SRI.

Zieky, M. J. (1995). A historical perspective on setting standards. In L. Crocker & M. J. Zieky (Eds.), *Joint conference on standard setting for large-scale assessments* (Vol. 2, pp. 1–38). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Educational Psychology, 20*, 79–88. https://doi.org/10.1016/j.pse.2014.11.003

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251. ps://doi.org/10.1111/j.1745-3984.1993.tb00425.x

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321–344. https://doi.org/10.1207/s15324818ame1004_2

Chapter 19
Advancing Human Assessment: A Synthesis Over Seven Decades

Randy E. Bennett and Matthias von Davier

This book has documented the history of ETS's contributions to educational research and policy analysis, psychology, and psychometrics. We close the volume with a brief synthesis in which we try to make more general meaning from the diverse directions that characterized almost 70 years of work.

Synthesizing the breadth and depth of the topics covered over that time period is not simple. One way to view the work is across time. Many of the book's chapters presented chronologies, allowing the reader to follow the path of a research stream over the years. Less evident from these separate chronologies was the extent to which multiple streams of work not only coexisted but sometimes interacted.

From its inception, ETS was rooted in Henry Chauncey's vision of describing individuals through broad assessment of their capabilities, helping them to grow and society to benefit (Elliot 2014). Chauncey's conception of broad assessment of capability required a diverse research agenda.

Following that vision, his research managers assembled an enormous range of staff expertise. Only through the assemblage of such expertise could one bring diverse perspectives and frameworks from many fields to a problem, leading to novel solutions.

In the following sections, we summarize some of the key research streams evident in different time periods, where each period corresponds to roughly a decade. Whereas the segmentation of these time periods is arbitrary, it does give a general

---

This work was conducted while M. von Davier was employed with Educational Testing Service.

R.E. Bennett (✉) • M. von Davier
Educational Testing Service, Princeton, NJ, USA
e-mail: rbennett@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_19

sense of the progression of topics across time.[1] Also somewhat arbitrary is the use of publication date as the primary determinant of placement into a particular decade. Although the work activity leading up to publication may well have occurred in the previous period, the result of that activity and the impact that it had was typically through its dissemination.

19.1 The Years 1948–1959

19.1.1 Psychometric and Statistical Methodology

As will be the case for every period, a very considerable amount of work centered on theory and on methodological development in psychometrics and statistics. With respect to the former, the release of Gulliksen's (1950) *Theory of Mental Tests* deserves special mention for its codification of classical test theory. But more forward looking was work to create a statistically grounded foundation for the analysis of test scores, a latent-trait theory (Lord 1952, 1953). This direction would later lead to the groundbreaking development of item response theory (IRT; Lord and Novick 1968), which became a well-established part of applied statistical research in domains well beyond education and is now an important building block of generalized modeling frameworks, which connect the item response functions of IRT with structural models (Carlson and von Davier, Chap. 5, this volume). Green's (1950a, b) work can be seen as an early example that has had continued impact not commonly recognized. His work pointed out how latent structure and latent-trait models are related to factor analysis, while at the same time placing latent-trait theory into the context of latent class models. Green's insights had profound impact, reemerging outside of ETS in the late 1980s (de Leeuw and Verhelst 1986; Follman 1988; Formann 1992; Heinen 1996) and, in more recent times, at ETS in work on generalized latent variable models (Haberman et al. 2008; Rijmen et al. 2014).

In addition to theoretical development, substantial effort was focused on methodological development for, among other purposes, the generation of engineering solutions to practical scale-linking problems. Examples include Karon and Cliff's (1957) proposal to smooth test-taker sample data before equating, a procedure used today by most testing programs that employ equipercentile equating (Dorans and Puhan, Chap. 4, this volume); Angoff's (1953) method for equating test forms by using a miniature version of the full test as an external anchor; and Levine's (1955) procedures for linear equating under the common-item, nonequivalent-population design.

[1] In most cases, citations included as examples of a work stream were selected based on their discussion in one of the book's chapters.

19.1.2 Validity and Validation

In the 2 years of ETS's beginning decade, the 1940s, and in the 1950s that followed, great emphasis was placed on predictive studies, particularly for success in higher education. Studies were conducted against first-semester performance (Frederiksen 1948) as well as 4-year academic criteria (French 1958). As Kane and Bridgeman (Chap. 16, this volume) noted, this emphasis was very much in keeping with conceptions of validity at the time, and it was, of course, important to evaluating the meaning and utility of scores produced by the new organization's operational testing programs. However, also getting attention were studies to facilitate trait interpretations of scores (French et al. 1952). These interpretations posited that response consistencies were the result of test-taker dispositions to behave in certain ways in response to certain tasks, dispositions that could be investigated through a variety of methods, including factor analysis. Finally, the compromising effects of construct-irrelevant influences, in particular those due to coaching, were already a clear concern (Dear 1958; French and Dear 1959).

19.1.3 Constructed-Response Formats and Performance Assessment

Notably, staff interests at this time were not restricted to multiple-choice tests because, as Bejar (Chap. 18, this volume) pointed out, the need to evaluate the value of additional methods was evident. Work on constructed-response formats and performance assessment was undertaken (Ryans and Frederiksen 1951), including development of the in-basket test (Fredericksen et al. 1957), subsequently used throughout the world for job selection, and a measure of the ability to formulate hypotheses as an indicator of scientific thinking (Frederiksen 1959). Research on direct writing assessment (e.g., through essay testing) was also well under way (Diederich 1957; Huddleston 1952; Torgerson and Green 1950).

19.1.4 Personal Qualities

Staff interests were not restricted to the verbal and quantitative abilities underlying ETS's major testing programs, the Scholastic Aptitude Test (the *SAT*® test) and the *GRE*® General Test. Rather, a broad investigative program on what might be termed *personal qualities* was initiated. Cognition, more generally defined, was one key interest, as evidenced by publication of the Kit of Selected Tests for Reference Aptitude and Achievement Factors (French 1954). The Kit was a compendium of marker assessments investigated with sufficient thoroughness to make it possible to use in factor analytic studies of cognition such that results could be more directly

compared across studies. Multiple reference measures were provided for each factor, including measures of abilities in the reasoning, memory, spatial, verbal, numeric, motor, mechanical, and ideational fluency domains.

In addition, substantial research targeted a wide variety of other human qualities. This research included personality traits, interests, social intelligence, motivation, leadership, level of aspiration and need for achievement, and response styles (acquiescence and social desirability), among other things (French 1948, 1956; Hills 1958; Jackson and Messick 1958; Melville and Frederiksen 1952; Nogee 1950; Ricciuti 1951).

19.2 The Years 1960–1969

19.2.1 Psychometric and Statistical Methodology

If nothing else, this period was notable for the further development of IRT (Lord and Novick 1968). That development is one of the major milestones of psychometric research. Although the organization made many important contributions to classical test theory, today psychometrics around the world mainly uses IRT-based methods, more recently in the form of generalized latent variable models. One of the important differences from classical approaches is that IRT properly grounds the treatment of categorical data in probability theory and statistics. The theory's modeling of how responses statistically relate to an underlying variable allows for the application of powerful methods for generalizing test results and evaluating the assumptions made. IRT-based item functions are the building blocks that link item responses to underlying explanatory models (Carlson and von Davier, Chap. 5, this volume). Leading up to and concurrent with the seminal volume *Statistical Theories of Mental Test Scores* (Lord and Novick 1968), Lord continued to make key contributions to the field (Lord 1965a, b, 1968a, b).

In addition to the preceding landmark developments, a second major achievement was the invention of confirmatory factor analysis by Karl Jöreskog (1965, 1967, 1969), a method for rigorously evaluating hypotheses about the latent structure underlying a measure or collection of measures. This invention would be generalized in the next decade and applied to the solution of a great variety of measurement and research problems.

19.2.2 Large-Scale Survey Assessments of Student and Adult Populations

In this period, ETS contributed to the design and conducted the analysis of the Equality of Educational Opportunity Study (Beaton and Barone, Chap. 8, this volume). Also of note was that, toward the end of the decade, ETS's long-standing program of longitudinal studies began with initiation of the Head Start Longitudinal Study (Anderson et al. 1968). This study followed a sample of children from before preschool enrollment through their experience in Head Start, in another preschool, or in no preschool program.

19.2.3 Validity and Validation

The 1960s saw continued interest in prediction studies (Schrader and Pitcher 1964), though noticeably less than in the prior period. The study of construct-irrelevant factors that had concentrated largely on coaching was less evident, with interest emerging in the phenomenon of test anxiety (French 1962). Of special note is that, due to the general awakening in the country over civil rights, ETS research staff began to focus on developing conceptions of equitable treatment of individuals and groups (Cleary 1968).

19.2.4 Constructed-Response Formats and Performance Assessment

The 1960s saw much investigation of new forms of assessment, including in-basket performance (Frederiksen 1962; L. B. Ward 1960), formulating-hypotheses tasks (Klein et al. 1969), and direct writing assessment. As described by Bejar (Chap. 18, this volume), writing assessment deserves special mention for the landmark study by Diederich et al. (1961) documenting that raters brought "schools of thought" to the evaluation of essays, thereby initiating interest in the investigation of rater cognition, or the mental processes underlying essay grading. A second landmark was the study by Godshalk et al. (1966) that resulted in the invention of holistic scoring.

19.2.5 Personal Qualities

The 1960s brought a very substantial increase to work in this area. The work on cognition produced the 1963 "Kit of Reference Tests for Cognitive Factors" (French et al. 1963), the successor to the 1954 "Kit." Much activity concerned the measurement of personality specifically, although a range of related topics was also investigated, including continued work on response styles (Damarin and Messick 1965; Jackson and Messick 1961; Messick 1967), the introduction into the social–psychological literature of the concept of prosocial (or altruistic) behavior (Bryan and Test 1967; Rosenhan 1969; Rosenhan and White 1967), and risk taking (Kogan and Doise 1969; Kogan and Wallach 1964; Wallach et al. 1962). Also of note is that this era saw the beginnings of ETS's work on cognitive styles (Gardner et al. 1960; Messick and Fritzky 1963; Messick and Kogan 1966). Finally, a research program on creativity began to emerge (Skager et al. 1965, 1966), including Kogan's studies of young children (Kogan and Morgan 1969; Wallach and Kogan 1965), a precursor to the extensive line of developmental research that would appear in the following decade.

19.2.6 Teacher and Teaching Quality

Although ETS had been administering the National Teachers Examination since the organization's inception, relatively little research had been conducted around the evaluation of teaching and teachers. The 1960s saw the beginnings of such research, with investigations of personality (Walberg 1966), values (Sprinthall and Beaton 1966), and approaches to the behavioral observation of teaching (Medley and Hill 1967).

19.3 The Years 1970–1979

19.3.1 Psychometric and Statistical Methodology

Causal inference was a major area of research in the field of statistics generally in this decade, and that activity included ETS. Rubin (1974b, 1976a, b, c, 1978) made fundamental contributions to the approach that allows for evaluating the extent to which differences observed in experiments can be attributed to effects of underlying variables.

More generally, causal inference as treated by Rubin can be understood as a missing-data and imputation problem. The estimation of quantities under incomplete-data conditions was a chief focus, as seen in work by Rubin (1974a, 1976a, b) and his collaborators (Dempster et al. 1977), who created the

expectation-maximization (EM) algorithm, which has become a standard analytical method used not only in estimating modern psychometric models but throughout the sciences. As of this writing, the Dempster et al. (1977) article had more than 45,000 citations in Google Scholar.

Also falling under causal inference was Rubin's work on matching. Matching was developed to reduce bias in causal inferences using data from nonrandomized studies. Rubin's (1974b, 1976a, b, c, 1979) work was central to evaluating and improving this methodology.

Besides landmark contributions to causal inference, continued development of IRT was taking place. Apart from another host of papers by Lord (1970, 1973, 1974a, b, 1975a, b, 1977), several applications of IRT were studied, including for linking test forms (Marco 1977; see also Carlson and von Davier, Chap. 5, this volume). In addition, visiting scholars made seminal contributions as well. Among these contributions were ones on testing the Rasch model as well as on bias in estimates (Andersen 1972, 1973), ideas later generalized by scholars elsewhere (Haberman 1977).

Finally, this period saw Karl Jöreskog and colleagues implement confirmatory factor analysis (CFA) in the LISREL computer program (Jöreskog and van Thillo 1972) and generalize CFA for the analysis of covariance structures (Jöreskog 1970), path analysis (Werts et al. 1973), simultaneous factor analysis in several populations (Jöreskog 1971), and the measurement of growth (Werts et al. 1972). Their inventions, particularly LISREL, continue to be used throughout the social sciences within the general framework of structural equation modeling to pose and evaluate psychometric, psychological, sociological, and econometric theories and the hypotheses they generate.

19.3.2 Large-Scale Survey Assessments of Student and Adult Populations

Worthy of note were two investigations, one of which was a continuation from the previous decade. That latter investigation, the Head Start Longitudinal Study, was documented in a series of program reports (Emmerich 1973; Shipman 1972; Ward 1973). Also conducted was the National Longitudinal Study of the High School Class of 1972 (Rock, Chap. 10, this volume).

19.3.3 Validity and Validation

In this period, conceptions of validity, and concerns for validation, were expanding. With respect to conceptions of validity, Messick's (1975) seminal paper "The Standard Problem: Meaning and Values in Measurement and Evaluation" called

attention to the importance of construct interpretations in educational measurement, a perspective largely missing from the field at that time. As to validation, concerns over the effects of coaching reemerged with research finding that two quantitative item types being considered for the SAT were susceptible to short-term preparation (Evans and Pike 1973), thus challenging the College Board's position on the existence of such effects. Concerns for validation also grew with respect to test fairness and bias, with continued development of conceptions and methods for investigating these issues (Linn 1973, 1976; Linn and Werts 1971).

19.3.4 Constructed-Response Formats and Performance Assessment

Relatively little attention was given to this area. An exception was continued investigation of the formulating-hypotheses item type (Evans and Frederiksen 1974; Ward et al. 1980).

19.3.5 Personal Qualities

The 1970s saw the continuation of a significant research program on personal qualities. With respect to cognition, the third version of the "Factor Kit" was released in 1976: the "Kit of Factor-Referenced Cognitive Tests" (Ekstrom et al. 1976). Work on other qualities continued, including on prosocial behavior (Rosenhan 1970, 1972) and risk taking (Kogan et al. 1972; Lamm and Kogan 1970; Zaleska and Kogan 1971). Of special note was the addition to the ETS staff of Herman Witkin and colleagues, who significantly extended the prior decade's work on cognitive styles (Witkin et al. 1974, 1977; Zoccolotti and Oltman 1978). Work on kinesthetic aftereffect (Baker et al. 1976, 1978, 1979) and creativity (Frederiksen and Ward 1978; Kogan and Pankove 1972; Ward et al. 1972) was also under way.

19.3.6 Human Development

The 1970s saw the advent of a large work stream that would extend over several decades. This work stream might be seen as a natural extension of Henry Chauncey's interest in human abilities, broadly conceived; that is, to understand human abilities, it made sense to study from where those abilities emanated. That stream, described in detail by Kogan et al. (Chap. 15, this volume), included research in many areas. In this period, it focused on infants and young children, encompassing their social development (Brooks and Lewis 1976; Lewis and Brooks-Gunn 1979), emotional

development (Lewis 1977; Lewis et al. 1978; Lewis and Rosenblum 1978), cognitive development (Freedle and Lewis 1977; Lewis 1977, 1978), and parental influences (Laosa 1978; McGillicuddy-DeLisi et al. 1979).

19.3.7 Educational Evaluation and Policy Analysis

One of the more notable characteristics of ETS research in this period was the emergence of educational evaluation, in good part due to an increase in policy makers' interest in appraising the effects of investments in educational interventions. This work, described by Ball (Chap. 11, this volume), entailed large-scale evaluations of television programs like *Sesame Street* and *The Electric Company* (Ball and Bogatz 1970, 1973) and early computer-based instructional systems like PLATO and TICCIT (Alderman 1978; Murphy 1977), as well as a wide range of smaller studies (Marco 1972; Murphy 1973). Some of the accumulated wisdom gained in this period was synthesized in two books, the *Encyclopedia of Educational Evaluation* (Anderson et al. 1975) and *The Profession and Practice of Program Evaluation* (Anderson and Ball 1978).

Alongside the intense evaluation activity was the beginning of a work stream on policy analysis (see Coley et al., Chap. 12, this volume). That beginning concentrated on education finance (Goertz 1978; Goertz and Moskowitz 1978).

19.3.8 Teacher and Teaching Quality

Rounding out the very noticeable expansion of research activity that characterized the 1970s were several lines of work on teachers and teaching. One line concentrated on evaluating the functioning of the National Teachers Examination (NTE; Quirk et al. 1973). A second line revolved around observing and analyzing teaching behavior (Quirk et al. 1971, 1975). This line included the Beginning Teacher Evaluation Study, the purpose of which was to identify teaching behaviors effective in promoting learning in reading and mathematics in elementary schools, a portion of which was conducted by ETS under contract to the California Commission for Teacher Preparation and Licensing. The study included extensive classroom observation and analysis of the relations among the observed behaviors, teacher characteristics, and student achievement (McDonald and Elias 1976; Sandoval 1976). The final line of research concerned college teaching (Baird 1973; Centra 1974).

19.4 The Years 1980–1989

19.4.1 Psychometric and Statistical Methodology

As was true for the 1970s, in this decade, ETS methodological innovation was notable for its far-ranging impact. Lord (1980) furthered the development and application of IRT, with particular attention to its use in addressing a wide variety of testing problems, among them parameter estimation, linking, evaluation of differential item functioning (DIF), and adaptive testing. Holland (1986, 1987), as well as Holland and Rubin (1983), continued the work on causal inference, further developing its philosophical and epistemological foundations, including exploration of a long-standing statistical paradox described by Lord (1967).[2] An edited volume, *Drawing Inferences From Self-Selected Samples* (Wainer 1986), collected work on these issues.

Rubin's work on matching, particularly propensity score matching, was a key activity through this decade. Rubin (1980a), as well as Rosenbaum and Rubin (1984, 1985), made important contributions to this methodology. These widely cited publications outlined approaches that are frequently used in scientific research when experimental manipulation is not possible.

Building on his research of the previous decade, Rubin (1980b, c) developed "multiple imputation," a statistical technique for dealing with nonresponse by generating random draws from the posterior distribution of a variable, given other variables. The multiple imputations methodology forms the underlying basis for several major group-score assessments (i.e., tests for which the focus of inference is on population, rather than individual, performance), including the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the Programme of International Assessment of Adult Competencies (PIAAC; Beaton and Barone, Chap. 8, this volume; Kirsch et al., Chap. 9, this volume).

Also of note was the emergence of DIF as an important methodological research focus. The standardization method (Dorans and Kulick 1986), and the more statistically grounded Mantel and Haenszel (1959) technique proposed by Holland and Thayer (1988), became stock approaches used by operational testing programs around the world for assessing item-level fairness. Finally, the research community working on DIF was brought together for an invited conference in 1989 at ETS.

Although there were a large number of observed-score equating studies in the 1980s, one development stands out in that it foreshadowed a line of research undertaken more than a decade later. The method of kernel equating was introduced by Holland and Thayer (1989) as a general procedure that combines smoothing,

[2] Lord's (1967) paradox refers to the situation, in observational studies, in which the statistical treatment of posttest scores by means of different corrections using pretest scores (i.e., regression vs. posttest minus pretest differences) can lead to apparent contradictions in results. This phenomenon is related to regression artifacts (D. T. Campbell and Kenny, 1999; Eriksson and Haggstrom, 2014).

modeling, and transforming score distributions. This combination of statistical procedures was intended to provide a flexible tool for observed-score equating in a nonequivalent-groups anchor-test design.

19.4.2 Large-Scale Survey Assessments of Student and Adult Populations

ETS was first awarded the contract for NAEP in 1983 after evaluating previous NAEP analytic procedures and releasing *A New Design for a New Era* (Messick et al. 1983). The award set the stage for advances in assessment design and psychometric methodology, including extensions of latent-trait models that employed covariates. These latent regression models used maximum likelihood methods to estimate population parameters from observed item responses without estimating individual ability parameters for test takers (Mislevy 1984, 1985). Many of the approaches developed for NAEP were later adopted by other national and international surveys, including the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS), PISA, and PIAAC. These surveys are either directly modeled on NAEP or are based on other surveys that were themselves NAEP's direct derivates.

The major design and analytic features shared by these surveys include (a) a balanced incomplete block design that allows broad coverage of content frameworks, (b) use of modern psychometric methods to link across the multiple test forms covering this content, (c) integration of cognitive tests and respondent background data using those psychometric methods, and (d) a focus on student (and adult) populations rather than on individuals as the targets of inference and reporting.

Two related developments should be mentioned. The chapters by Kirsch et al. (Chap. 9, this volume) and Rock (Chap. 10, this volume) presented in more detail work on the 1984 Young Adult Literacy Study (YALS) and the 1988 National Educational Longnitudinal Study, respectively. These studies also use multiple test forms and advanced psychometric methods based on IRT. Moreover, YALS was the first to apply a multidimensional item response model (Kirsch and Jungeblut 1986).

19.4.3 Validity and Validation

The 1980s saw the culmination of Messick's landmark unified model (Messick 1989), which framed validity as a unitary concept. The highlight of the period, Messick's chapter in *Educational Measurement*, brought together the major strands of validity theory, significantly influencing conceptualization and practice throughout the field.

Also in this period, research on coaching burgeoned in response to widespread public and institutional user concerns (see Powers, Chap. 17, this volume). Notable was publication of *The Effectiveness of Coaching for the SAT: Review and Reanalysis of Research From the Fifties to the FTC* (Messick 1980), though many other studies were also released (Alderman and Powers 1980; Messick 1982; Powers 1985; Powers and Swinton 1984; Swinton and Powers 1983). Other sources of construct-irrelevant variance were investigated, particularly test anxiety (Powers 1988). Finally, conceptions of fairness became broader still, motivated by concerns over the flagging of scores from admissions tests that were administered under nonstandard conditions to students with disabilities; these concerns had been raised most prominently by a National Academy of Sciences panel (Sherman and Robinson 1982). Most pertinent was the 4-year program of research on the meaning and use of such test scores for the SAT and GRE General Test that was initiated in response to the panel's report. Results were summarized in the volume *Testing Handicapped People* by Willingham et al. (1988).

19.4.4 Constructed-Response Formats and Performance Assessment

Several key publications highlighted this period. Frederiksen's (1984) *American Psychologist* article "The Real Test Bias: Influences of Testing on Teaching and Learning" made the argument for the use of response formats in assessment that more closely approximated the processes and outcomes important for success in academic and work environments. This classic article anticipated the K–12 performance assessment movement of the 1990s and its 2010 resurgence in the Common Core Assessments. Also noteworthy were Breland's (1983) review showing the incremental predictive value of essay tasks over multiple-choice measures at the postsecondary level and his comprehensive study of the psychometric characteristics of such tasks (Breland et al. 1987). The Breland et al. volume included analyses of rater agreement, generalizability, and dimensionality. Finally, while research continued on the formulating-hypotheses item type (Ward et al. 1980), the investigation of portfolios also emerged (Camp 1985).

19.4.5 Personal Qualities

Although investigation of cognitive style continued in this period (Goodenough et al. 1987; Messick 1987; Witkin and Goodenough 1981), the death of Herman Witkin in 1979 removed its intellectual leader and champion, contributing to its decline. This decline coincided with a drop in attention to personal qualities research more generally, following a shift in ETS management priorities from the very clear

think tank orientation of the 1960s and 1970s to a greater focus on research to assist existing testing programs and the creation of new ones. That focus remained centered largely on traditional academic abilities, though limited research proceeded on creativity (Baird and Knapp 1981; Ward et al. 1980).

19.4.6 Human Development

Whereas the research on personal qualities noticeably declined, the work on human development remained vibrant, at least through the early part of this period, in large part due to the availability of external funding and staff members highly skilled at attracting it. With a change in management focus, the reassignment of some developmental staff to other work, and the subsequent departure of the highly prolific Michael Lewis, interest began to subside. Still, this period saw a considerable amount and diversity of research covering social development (Brooks-Gunn and Lewis 1981; Lewis and Feiring 1982), emotional development (Feinman and Lewis 1983; Lewis and Michalson 1982), cognitive development (Lewis and Brooks-Gunn 1981a, b; Sigel 1982), sexual development (Brooks-Gunn 1984; Brooks-Gunn and Warren 1988), development of Chicano children (Laosa 1980a, 1984), teenage motherhood (Furstenberg et al. 1987), perinatal influences (Brooks-Gunn and Hearn 1982), parental influences (Brody et al. 1986; Laosa 1980b), atypical development (Brinker and Lewis 1982; Brooks-Gunn and Lewis 1982), and interventions for vulnerable children (Brooks-Gunn et al. 1988; Lee et al. 1988).

19.4.7 Educational Evaluation and Policy Analysis

As with personal qualities, the evaluation of educational programs began to decline during this period. In contrast to the work on personal qualities, evaluation activities had been almost entirely funded through outside grants and contracts, which diminished considerably in the 1980s. In addition, the organization's most prominent evaluator, Samuel Ball, departed to take an academic appointment in his native Australia. The work that remained investigated the effects of instructional software like the IBM Writing to Read program (Murphy and Appel 1984), educational television (Murphy 1988), alternative higher education programs (Centra and Barrows 1982), professional training (Campbell et al. 1982), and the educational integration of students with severe disabilities (Brinker and Thorpe 1984).

Whereas funding for evaluation was in decline, support for policy analysis grew. Among other things, this work covered finance (Berke et al. 1984), teacher policy (Goertz et al. 1984), education reform (Goertz 1989), gender equity (Lockheed 1985), and access to and participation in graduate education (Clewell 1987).

19.4.8 Teacher and Teaching Quality

As with program evaluation, the departure of key staff during this period resulted in diminished activity, with only limited attention given to the three dominant lines of research of the previous decade: functioning of the NTE (Rosner and Howey 1982), classroom observation (Medley and Coker 1987; Medley et al. 1981), and college teaching (Centra 1983). Of particular note was Centra and Potter's (1980) article "School and Teacher Effects: An Interrelational Model," which proposed an early structural model for evaluating input and context variables in relation to achievement.

19.5 The Years 1990–1999

19.5.1 Psychometric and Statistical Methodology

DIF continued to be an important methodological research focus. In the early part of the period, an edited volume, *Differential Item Functioning*, was released based on the 1989 DIF conference (Holland and Wainer 1993). Among other things, the volume included research on the Mantel–Haenszel (1959) procedure. Other publications, including on the standardization method, have had continued impact on practice (Dorans and Holland 1993; Dorans et al. 1992). Finally, of note were studies that placed DIF into model-based frameworks. The use of mixture models (Gitomer and Yamamoto 1991; Mislevy and Verhelst 1990; Yamamoto and Everson 1997), for example, illustrated how to relax invariance assumptions and test DIF in generalized versions of item response models.

Among the notable methodological book publications of this period was *Computer Adaptive Testing: A Primer*, edited by Wainer et al. (1990). This volume contained several chapters by ETS staff members and their colleagues.

Also worthy of mention was research on extended IRT models, which resulted in several major developments. Among these developments were the generalized partial credit model (Muraki 1992), extensions of mixture IRT models (Bennett et al. 1991; Gitomer and Yamamoto 1991; Yamamoto and Everson 1997), and models that were foundational for subsequent generalized modeling frameworks. Several chapters in the edited volume *Test Theory for a New Generation of Tests* (Frederiksen et al. 1993) described developments around these extended IRT models.

19.5.2 Large-Scale Survey Assessments of Student and Adult Populations

NAEP entered its second decade with the new design and analysis methodology introduced by ETS. Articles describing these methodological innovations were published in a special issue of the *Journal of Educational Statistics* (Mislevy et al. 1992b; Yamamoto and Mazzeo 1992). Many of these articles remain standard references, used as a basis for extending the methods and procedures of group-score assessments. In addition, Mislevy (1991, 1993a, b) continued work on related issues.

A significant extension to the large-scale assessment work was a partnership with Statistics Canada that resulted in development of the International Adult Literacy Survey (IALS). IALS collected data in 23 countries or regions of the world, 7 in 1994 and an additional 16 in 1996 and 1998 (Kirsch et al., Chap. 9, this volume). Also in this period, ETS research staff helped the International Association for the Evaluation of Educational Achievement (IEA) move the TIMSS 1995 and 1999 assessments to a more general IRT model, later described by Yamamoto and Kulick (2002). Finally, this period saw the beginning of the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K), which followed students through the eighth grade (Rock, Chap. 10, this volume).

19.5.3 Validity and Validation

Following the focus on constructs advocated by Messick's (1989) chapter, the 1990s saw a shift in thinking that resulted in concerted attempts to ground assessment design in domain theory, particularly in domains in which design had been previously driven by content frameworks. Such theories often offered a deeper and clearer description of the cognitive components that made for domain proficiency and the relationships among the components. A grounding in cognitive-domain theory offered special advantages for highly interactive assessments like simulations because of the expense involved in their development, which increased dramatically without the guidance provided by theory for task creation and scoring. From Messick (1994a), and from work on an intelligent tutoring system that combined domain theory with rigorous probability models (Gitomer et al. 1994), the foundations of evidence-centered design (ECD) emerged (Mislevy 1994, 1996). ECD, a methodology for rigorously reasoning from assessment claims to task development, and from item responses back to claims, is now used throughout the educational assessment community as a means of creating a stronger validity argument a priori.

During this same period, other investigators explored how to estimate predictive validity coefficients by taking into account differences in grading standards across college courses (Ramist et al. 1994). Finally, fairness for population groups remained

in focus, with continued attention to admissions testing for students with disabilities (Bennett 1999) and release of the book *Gender and Fair Assessment* by Willingham and Cole (1997), which comprehensively examined the test performance of males and females to identify potential sources of unfairness and possible solutions.

19.5.4 Constructed-Response Formats and Performance Assessment

At both the K–12 and postsecondary levels, interest in moving beyond multiple-choice measures was widespread. ETS work reflected that interest and, in turn, contributed to it. Highlights included Messick's (1994a) paper on evidence and consequences in the validation of performance assessments, which provided part of the conceptual basis for the invention of ECD, and publication of the book *Construction Versus Choice in Cognitive Measurement* (Bennett and Ward 1993), framing the breadth of issues implicated in the use of non-multiple-choice formats.

In this period, many aspects of the functioning of constructed-response formats were investigated, including construct equivalence (Bennett et al. 1991; Bridgeman 1992), population invariance (Breland et al. 1994; Bridgeman and Lewis 1994), and effects of allowing test takers choice in task selection (Powers and Bennett 1999). Work covered a variety of presentation and response formats, including formulating hypotheses (Bennett and Rock 1995), portfolios (Camp 1993; LeMahieu et al. 1995), and simulations for occupational and professional assessment (Steinberg and Gitomer 1996).

Appearing in this decade were ETS's first attempts at automated scoring, including of computer science subroutines (Braun et al. 1990), architectural designs (Bejar 1991), mathematical step-by-step solutions and expressions (Bennett et al. 1997; Sebrechts et al. 1991), short-text responses (Kaplan 1992), and essays (Kaplan et al. 1995). By the middle of the decade, the work on scoring architectural designs had been implemented operationally as part of the National Council of Architectural Registration Board's Architect Registration Examination (Bejar and Braun 1999). Also introduced at the end of the decade into the Graduate Management Admission Test was the *e-rater*® automated scoring engine, an approach to automated essay scoring (Burstein et al. 1998). The e-rater scoring engine continues to be used operationally for the GRE General Test Analytical Writing Assessment, the *TOEFL*® test, and other examinations.

19.5.5 Personal Qualities

Interest in this area had been in decline since the 1980s. The 1990s brought an end to the cognitive styles research, with only a few publications released (Messick 1994b, 1996). Some research on creativity continued (Bennett and Rock 1995; Enright et al. 1998).

19.5.6 Human Development

As noted, work in this area also began to decline in the 1980s. The 1990s saw interest diminish further with the departure of Jeanne Brooks-Gunn, whose extensive publications covered an enormous substantive range. Still, a significant amount of research was completed, including on parental influences and beliefs (Sigel 1992), representational competence (Sigel 1999), the distancing model (Sigel 1993), the development of Chicano children (Laosa 1990), and adolescent sexual, emotional, and social development (Brooks-Gunn 1990).

19.5.7 Education Policy Analysis

This period saw the continuation of a vibrant program of policy studies. Multiple areas were targeted, including finance (Barton et al. 1991), teacher policy (Bruschi and Coley 1999), education reform (Barton and Coley 1990), education technology (Coley et al. 1997), gender equity (Clewell et al. 1992), education and the economy (Carnevale 1996; Carnevale and DesRochers 1997), and access to and participation in graduate education (Ekstrom et al. 1991; Nettles 1990).

19.5.8 Teacher and Teaching Quality

In this period, a resurgence of interest occurred due to the need to build the foundation for the *PRAXIS*® program, which replaced the NTE. An extensive series of surveys, job analyses, and related studies was conducted to understand the knowledge, skills, and abilities required for newly licensed teachers (Reynolds et al. 1992; Tannenbaum 1992; Tannenbaum and Rosenfeld 1994). As in past decades, work was done on classroom performance (Danielson and Dwyer 1995; Powers 1992), some of which supplied the initial foundation for the widely used *Framework for Teaching Evaluation Instrument* (Danielson 2013).

19.6 The Years 2000–2009

19.6.1 Psychometric and Statistical Methodology

The first decade of the current century saw increased application of Bayesian methods in psychometric research, in which staff members continued ETS's tradition of integrating advances in statistics with educational and psychological measurement. Among the applications were posterior predictive checks (Sinharay 2003), a method not unlike the frequentist resampling and resimulation studied in the late 1990s (M. von Davier 1997), as well as the use of Bayesian networks to specify complex measurement models (Mislevy et al. 2000). Markov chain Monte Carlo methods were employed to explore the comprehensive estimation of measurement and structural models in modern IRT (Johnson and Jenkins 2005) but, because of their computational requirements, currently remain limited to small- to medium-sized applications.

Alternatives to these computationally demanding methods were considered to enable the estimation of high-dimensional models, including empirical Bayes methods and approaches that utilized Monte Carlo integration, such as the stochastic EM algorithm (M. von Davier and Sinharay 2007).

These studies were aimed at supporting the use of explanatory IRT applications taking the form of a latent regression that includes predictive background variables in the structural model. Models of this type are used in the NAEP, PISA, PIAAC, TIMSS, and PIRLS assessments, which ETS directly or indirectly supported. Sinharay and von Davier (2005) also presented extensions of the basic numerical integration approach to data having more dimensions. Similar to Johnson and Jenkins (2005), who proposed a Bayesian hierarchical model for the latent regression, Li et al. (2009) examined the use of hierarchical linear (or multilevel) extensions of the latent regression approach.

The kernel equating procedures proposed earlier by Holland and Thayer (1989; also Holland et al. 1989) were extended and designs for potential applications were described in *The Kernel Method of Test Equating* by A. A. von Davier, Holland, and Thayer (2004). The book's framework for observed-score equating encapsulates several well-known classical methods as special cases, from linear to equipercentile approaches.

A major reference work was released, titled *Handbook of Statistics: Vol. 26. Psychometrics* and edited by Rao and Sinharay (2006). This volume contained close to 1200 pages and 34 chapters reviewing state-of-the-art psychometric modeling. Sixteen of the volume's chapters were contributed by current or former ETS staff members.

The need to describe test-taker strengths and weaknesses has long motivated the reporting of subscores on tests that were primarily designed to provide a single score. Haberman (2008) presented the concept of proportional reduction of mean squared errors, which allows an evaluation of whether subscores are technically defensible. This straightforward extension of classical test theory derives from a

formula introduced by Kelley (1927) and provides a tool to check whether a subscore is reliable enough to stand on its own or whether the true score of the subscore under consideration would be better represented by the observed total score. (Multidimensional IRT was subsequently applied to this issue by Haberman and Sinharay 2010, using the same underlying argument.)

Also for purposes of better describing test-taker strengths and weaknesses, generalized latent variable models were explored, but with the intention of application to tests designed to measure multiple dimensions. Apart from the work on Bayesian networks (Mislevy and Levy 2007; Mislevy et al. 2003), there were significant extensions of approaches tracing back to the latent class models of earlier decades (Haberman 1988) and to the rule space model (Tatsuoka 1983). Among these extensions were developments around the reparameterized unified model (DiBello et al. 2006), which was shown to partially alleviate the identification issues of the earlier unified model, as well as around the general diagnostic model (GDM; M. von Davier 2008a). The GDM was shown to include many standard and extended IRT models, as well as several diagnostic models, as special cases (M. von Davier 2008a, b). The GDM has been successfully applied to the *TOEFL iBT*® test, PISA, NAEP, and PIRLS data in this as well as in the subsequent decade (M. von Davier 2008a; Oliveri and von Davier 2011, 2014; Xu and von Davier 2008). Other approaches later developed outside of ETS, such as the log-linear cognitive diagnostic model (LCDM; Henson et al. 2009), can be directly traced to the GDM (e.g., Rupp et al. 2010) and have been shown to be a special case of the GDM (M. von Davier 2014).

19.6.2 Large-Scale Survey Assessments of Student and Adult Populations

As described by Rock (Chap. 10, this volume), the Early Childhood Longitudinal Study continued through much of this decade, with the last data collection in the eighth grade, taking place in 2007. Also, recent developments in the statistical procedures used in NAEP were summarized and future directions described (M. von Davier et al. 2006).

A notable milestone was the Adult Literacy and Lifeskills (ALL) assessment, conducted in 2003 and 2006–2008 (Kirsch et al., Chap. 9, this volume). ALL was a household-based, international comparative study designed to provide participating countries with information about the literacy and numeracy skills of their adult populations. To accomplish this goal, ALL used nationally representative samples of 16- to 65-year-olds.

In this decade, ETS staff members completed a multicountry feasibility study for PISA of computer-based testing in multiple languages (Lennon, Kirsch, von Davier, Wagner, and Yamamoto 2003) and a report on linking and linking stability (Mazzeo and von Davier 2008).

Finally, in 2006, ETS and IEA established the IEA/ETS research institute (IERI), which promotes research on large-scale international skill surveys, publishes a journal, and provides training around the world through workshops on statistical and psychometric topics (Wagemaker and Kirsch 2008).

19.6.3 Validity and Validation

In the 2000s, Mislevy and colleagues elaborated the theory and generated additional prototypic applications of ECD (Mislevy et al. 2003, 2006), including proposing extensions of the methodology to enhance accessibility for individuals from special populations (Hansen and Mislevy 2006). Part of the motivation behind ECD was the need to more deeply understand the constructs to be measured and to use that understanding for assessment design. In keeping with that motivation, the beginning of this period saw the release of key publications detailing construct theory for achievement domains, which feed into the domain analysis and modeling aspects of ECD. Those publications concentrated on elaborating the construct of communicative competence for the TOEFL computer-based test (CBT), comprising listening, speaking, writing, and reading (Bejar et al. 2000; Butler et al. 2000; Cumming et al. 2000; Enright et al. 2000). Toward the end of the period, the Cognitively Based Assessment of, for, and as Learning (*CBAL*®) initiative (Bennett and Gitomer 2009) was launched. This initiative took a similar approach to construct definition as TOEFL CBT but, in CBAL's case, to the definition of English language arts and mathematics constructs for elementary and secondary education.

At the same time, the communication of predictive validity results for postsecondary admissions tests was improved. Building upon earlier work, Bridgeman and colleagues showed how the percentage of students who achieved a given grade point average increased as a function of score level, a more easily understood depiction than the traditional validity coefficient (Bridgeman et al. 2008). Also advanced was the research stream on test anxiety, one of several potential sources of irrelevant variance (Powers 2001).

Notable too was the increased attention given students from special populations. For students with disabilities, two research lines dominated, one related to testing and validation concerns that included but went beyond the postsecondary admissions focus of the 1980s and 1990s (Ekstrom and Smith 2002; Laitusis et al. 2002), and the second on accessibility (Hansen et al. 2004; Hansen and Mislevy 2006; Hansen et al. 2005). For English learners, topics covered accessibility (Hansen and Mislevy 2006; Wolf and Leon 2009), accommodations (Young and King 2008), validity frameworks and assessment guidelines (Pitoniak et al. 2009; Young 2009), and instrument and item functioning (Martiniello 2009; Young et al. 2008).

19.6.4 Constructed-Response Formats and Performance Assessment

Using ECD, several significant computer-based assessment prototypes were developed, including for NAEP (Bennett et al. 2007) and for occupational and professional assessment (Mislevy et al. 2002). The NAEP Technology-Rich Environments project was significant because assessment tasks involving computer simulations were administered to nationally representative samples of students and because it included an analysis of students' solution processes. This study was followed by NAEP's first operational technology-based component, the Interactive Computer Tasks, as part of the 2009 science assessment (U.S. Department of Education, n.d.-a). Also of note was the emergence of research on games and assessment (Shute et al. 2008, 2009).

With the presentation of constructed-response formats on computer came added impetus to investigate the effect of computer familiarity on performance. That issue was explored for essay tasks in NAEP (Horkay et al. 2006) as well as for the entry of complex expressions in mathematical reasoning items (Gallagher et al. 2002).

Finally, attention to automated scoring increased considerably. Streams of research on essay scoring and short-text scoring expanded (Attali and Burstein 2006; Leacock and Chodorow 2003; Powers et al. 2002; Quinlan et al. 2009), a new line on speech scoring was added (Zechner et al. 2007, 2009), and publications were released on the grading of graphs and mathematical expressions (Bennett et al. 2000).

19.6.5 Personal Qualities

Although it almost disappeared in the 1990s, ETS's interest in this topic reemerged following from the popularization of so-called noncognitive constructs in education, the workplace, and society at large (Goleman 1995). Two highly visible topics accounted for a significant portion of the research effort, one being emotional intelligence (MacCann and Roberts 2008; MacCann et al. 2008; Roberts et al. 2006) and the other stereotype threat (Stricker and Bejar 2004; Stricker and Ward 2004), the notion that concern about a negative belief as to the ability of one's demographic group might adversely affect test performance.

19.6.6 Human Development

With the death of Irving Sigel in 2006, the multidecade history of contributions to this area ended. Before his death, however, Sigel continued to write actively on the distancing model, representation, parental beliefs, and the relationship between

research and practice generally (Sigel 2000, 2006). Notable in this closing period was publication of his coedited *Child Psychology in Practice*, volume 4 of the *Handbook of Child Psychology* (Renninger and Sigel 2006).

19.6.7 Education Policy Analysis

Work in this area increased considerably. Several topics stood out for the attention given them. In elementary and secondary education, the achievement gap (Barton 2003), gender equity (Coley 2001), the role of the family (Barton and Coley 2007), and access to advanced course work in high school (Handwerk et al. 2008) were each examined. In teacher policy and practice, staff examined approaches to teacher preparation (Wang et al. 2003) and the quality of the teaching force (Gitomer 2007b).

With respect to postsecondary populations, new analyses were conducted of data from the adult literacy surveys (Rudd et al. 2004; Sum et al. 2002), and access to graduate education was studied (Nettles and Millett 2006). A series of publications by Carnevale and colleagues investigated the economic value of education and its equitable distribution (Carnevale and Fry 2001, 2002; Carnevale and Rose 2000). Among the many policy reports released, perhaps the highlight was *America's Perfect Storm* (Kirsch et al. 2007), which wove labor market trends, demographics, and student achievement into a social and economic forecast that received international media attention.

19.6.8 Teacher and Teaching Quality

Notable in this period were several lines of research. One centered on the functioning and impact of the certification assessments created by ETS for the National Board of Professional Teaching Standards (Gitomer 2007a; Myford and Engelhard 2001), which included the rating of video-recorded classroom performances. A second line more generally explored approaches for the evaluation of teacher effectiveness and teaching quality (Gitomer 2009; Goe et al. 2008; Goe and Croft 2009) as well as the link between teaching quality and student outcomes (Goe 2007). Deserving special mention was Braun's (2005) report "Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models," which called attention to the problems with this approach. Finally, a third work stream targeted professional development, including enhancing teachers' formative assessment practices (Thompson and Goe 2009; Wylie et al. 2009).

19.7 The Years 2010–2016

19.7.1 Psychometric and Statistical Methodology

Advances in computation have historically been an important driver of psychometric developments. In this period, staff members continued to create software packages, particularly for complex multidimensional analyses. One example was software for the operational use of multidimensional item response theory (MIRT) for simultaneous linking of multiple assessments (Haberman 2010). Another example was software for the operational use of the multidimensional discrete latent-trait model for IRT (and MIRT) calibration and linking (M. von Davier and Rost 2016). This software is used extensively for PIAAC and PISA.

Whereas software creation has constituted a continued line of activity, research on how to reduce computational burden has also been actively pursued. Of note in this decade was the use of graphical modeling frameworks to reduce the calculations required for complex multidimensional estimation. Rijmen (2010) as well as Rijmen et al. (2014) showed how these advances can be applied in large-scale testing applications, producing research software for that purpose. On a parallel track, von Davier (2016) described the use of all computational cores of a workstation or server to solve measurement problems in many dimensions more efficiently and to analyze the very large data sets coming from online testing and large-scale assessments of national or international populations.

In the same way as advances in computing have spurred methodological innovation, those computing advances have made the use of new item response types more feasible (Bejar, Chap. 18, this volume). Such response types have, in turn, made new analytic approaches necessary. Research has examined psychometric models and latent-trait estimation for items with multiple correct choices, self-reports using anchoring vignettes, data represented as multinomial choice trees, and responses collected from interactive and simulation tasks (Anguiano-Carrasco et al. 2015; Khorramdel and von Davier 2014), in the last case including analysis of response time and solution process.

Notable methodological publications collected in edited volumes in this period covered linking (von Davier 2011), computerized multistage testing (Yan et al. 2014), and international large-scale assessment methodology (Rutkowski et al. 2013). In addition, several contributions appeared by ETS authors in a three-volume handbook on IRT (Haberman 2016; von Davier and Rost 2016). Chapters by other researchers detail methods and statistical tools explored while those individuals were at ETS (e.g., Casabianca and Junker 2016; Moses 2016; Sinharay 2016).

19.7.2 Large-Scale Survey Assessments of Student and Adult Populations

In this second decade of the twenty-first century, the work of many research staff members was shaped by the move to computer-based, large-scale assessment. ETS became the main contractor for the design, assessment development, analysis, and project management of both PIAAC and PISA. PIAAC was fielded in 2012 as a multistage adaptive test (Chen et al. 2014b). In contrast, PISA 2015 was administered as a linear test with three core domains (science, mathematics, and reading), one innovative assessment domain (collaborative problem solving), and one optional domain (financial literacy).

NAEP also fielded computer-based assessments in traditional content domains and in domains that would not be suitable for paper-and-pencil administration. Remarkable were the delivery of the 2011 NAEP writing assessment on computer (U.S. Department of Education, n.d.-b) and the 2014 Technology and Engineering Literacy assessment (U.S. Department of Education, n.d.-c). The latter assessment contained highly interactive simulation tasks involving the design of bicycle lanes and the diagnosis of faults in a water pump. A large pilot study exploring multistage adaptive testing was also carried out (Oranje and Ye 2013) as part of the transition of all NAEP assessments to administration on computers.

Finally, ETS received the contract for PISA 2018, which will also entail the use of computer-based assessments in both traditional and nontraditional domains.

19.7.3 Validity and Validation

The work on construct theory in achievement domains for elementary and secondary education that was begun in the prior decade continued with publications in the English language arts (Bennett et al. 2016; Deane et al. 2015; Deane and Song 2015; Sparks and Deane 2015), mathematics (Arieli-Attali and Cayton-Hodges 2014; Graf 2009), and science (Liu et al. 2013). These publications detailed the CBAL competency, or domain, models and their associated learning progressions, that is, the pathways most students might be expected to take toward domain competency. Also significant was the Reading for Understanding project, which reformulated and exemplified the construct of reading comprehension for the digital age (Sabatini and O'Reilly 2013). Finally, a competency model was released for teaching (Sykes and Wilson 2015), intended to lay the foundation for a next generation of teacher licensure assessment.

In addition to domain modeling, ETS's work in validity theory was extended in several directions. The first direction was through further development of ECD, in particular its application to educational games (Mislevy et al. 2014). A second direction resulted from the arrival of Michael Kane, whose work on the argument-based approach added to the research program very substantially (Kane 2011, 2012,

2016). Finally, fairness and validity were combined in a common framework by Xi (2010).

Concerns for validity and fairness continued to motivate a wide-ranging research program directed at students from special populations. For those with disabilities, topics included accessibility (Hansen et al. 2012; Stone et al. 2016), accommodations (Cook et al. 2010), instrument and item functioning (Buzick and Stone 2011; Steinberg et al. 2011), computer-adaptive testing (Stone et al. 2013; Stone and Davey 2011), automated versus human essay scoring (Buzick et al. 2016), and the measurement of growth (Buzick and Laitusis 2010a, b). For English learners, topics covered accessibility (Guzman-Orth et al. 2016; Young et al. 2014), accommodations (Wolf et al. 2012a, b), instrument functioning (Gu et al. 2015; Young et al. 2010), test use (Lopez et al. 2016; Wolf and Farnsworth 2014; Wolf and Faulkner-Bond 2016), and the conceptualization of English learner proficiency assessment systems (Hauck et al. 2016; Wolf et al. 2016).

19.7.4 Constructed-Response Formats and Performance Assessment

As a consequence of growing interest in education, the work on games and assessment that first appeared at the end of the previous decade dramatically increased (Mislevy et al. 2012, 2014, 2016; Zapata-Rivera and Bauer 2012).

Work on automated scoring also grew substantially. The focus remained on response types from previous periods, such as essay scoring (Deane 2013a, b), short answer scoring (Heilman and Madnani 2012), speech scoring (Bhat and Yoon 2015; Wang et al. 2013), and mathematical responses (Fife 2013). However, important new lines of work were added. One such line, made possible by computer-based assessment, was the analysis of keystroke logs generated by students as they responded to essays, simulations, and other performance tasks (Deane and Zhang 2015; He and von Davier 2015, 2016; Zhang and Deane 2015). This analysis began to open a window into the processes used by students in problem solving. A second line, also made possible by advances in technology, was conversation-based assessment, in which test takers interact with avatars (Zapata-Rivera et al. 2014). Finally, a work stream was initiated on "multimodal assessment," incorporating analysis of test-taker speech, facial expression, or other behaviors (Chen et al. 2014a, c).

19.7.5 Personal Qualities

While work on emotional intelligence (MacCann et al. 2011; MacCann et al. 2010; Roberts et al. 2010), and stereotype threat (Stricker and Rock 2015) continued, this period saw a significant broadening to a variety of noncognitive constructs and their

applications. Research and product development were undertaken in education (Burrus et al. 2011; Lipnevich and Roberts 2012; Oliveri and Ezzo 2014) as well as for the workforce (Burrus et al. 2013; Naemi et al. 2014).

19.7.6 Education Policy Analysis

Although the investigation of economics and education had diminished due to the departure of Carnevale and his colleagues, attention to a wide range of policy problems continued. Those problems related to graduate education (Wendler et al. 2010), minority representation in teaching (Nettles et al. 2011), developing and implementing teacher evaluation systems (Goe et al. 2011), testing at the pre-K level (Ackerman and Coley 2012), achievement gaps in elementary and secondary education (Barton and Coley 2010), and parents opting their children out of state assessment (Bennett 2016).

A highlight of this period was the release of two publications from the ETS Opportunity Project. The publications, "Choosing Our Future: A Story of Opportunity in America" (Kirsch et al. 2016) and "The Dynamics of Opportunity in America" (Kirsch and Braun 2016), comprehensively analyzed and directed attention toward issues of equality, economics, and education in the United States.

19.7.7 Teacher and Teaching Quality

An active and diverse program of investigation continued. Support was provided for testing programs, including an extensive series of job analyses for revising PRAXIS program assessments (Robustelli 2010) as well as work toward the development of new assessments (Phelps and Howell 2016; Sykes and Wilson 2015). The general topic of teacher evaluation remained a constant focus (Gitomer and Bell 2013; Goe 2013; Turkan and Buzick 2016), including continued investigation into implementing it through classroom observation (Casabianca et al. 2013; Lockwood et al. 2015; Mihaly and McCaffrey 2014) and value-added modeling (Buzick and Jones 2015; McCaffrey 2013; McCaffrey et al. 2014). Researchers also explored the impact of teacher characteristics and teaching practices on student achievement (Liu et al. 2010), the effects of professional development on teacher knowledge (Bell et al. 2010), and the connection between teacher evaluation and professional learning (Goe et al. 2012). One highlight of the period was release of the fifth edition of AERA's *Handbook of Research on Teaching* (Gitomer and Bell 2016), a comprehensive reference for the field. A second highlight was *How Teachers Teach: Mapping the Terrain of Practice* (Sykes and Wilson 2015), which, as noted earlier, laid out a conceptualization of teaching in the form of a competency model.

19.8 Discussion

As the previous sections might suggest, the history of ETS research is marked by both constancy and changes in focus. The constancy can be seen in persistent attention to problems at the core of educational and psychological measurement. Those problems have centered on developing and improving the psychometric and statistical methodology that helps connect observations to inferences about individuals, groups, and institutions. In addition, the problems have centered on evaluating those inferences—that is, the theory, methodology, and practice of validation.

The changes in focus across time have occurred both within these two persistently pursued areas and among those areas outside of the measurement core. For example, Kane and Bridgeman (Chap. 16, this volume) documented in detail the progression that has characterized ETS's validity research, and multiple chapters did the same for the work on psychometrics and statistics. In any event, the emphasis given these core areas remained strong throughout ETS's history.

As noted, other areas experienced more obvious peaks and valleys. Several of these areas did not emerge as significant research programs in their own right until considerably after ETS was established. That characterization would be largely true, for example, of human development (beginning in the 1970s), educational evaluation (1970s), large-scale assessment/adult literacy/longitudinal studies (1970s), and policy analysis (1980s), although there were often isolated activities that preceded these dates. Once an area emerged, it did not necessarily persist, the best examples being educational evaluation, which spanned the 1970s to 1980s, and human development, which began at a similar time point, declined through the late 1980s and 1990s, and reached its denouement in the 2000s.

Still other areas rose, fell, and rose again. Starting with the founding of ETS, work on personal qualities thrived for three decades, all but disappeared in the 1980s and 1990s, and returned by the 2000s close to its past levels, but this time with the added focus of product development. The work on constructed-response formats and performance assessment also began early on and appeared to go dormant in the 1970s, only to return in the 1980s. In the 1990s, the emphasis shifted from a focus on paper-and-pencil measurement to presentation and scoring by computer.

What drove the constancy and change over the decades? The dynamics were most likely due to a complex interaction among several factors. One factor was certainly the influence of the external environment, including funding, federal education policy, public opinion, and the research occurring in the field. That environment, in turn, affected (and was affected by) the areas of interest and expertise of those on staff who, themselves, had impact on research directions. Finally the interests of the organization's management were affected by the external environment and, in turn, motivated actions that helped determine the staff composition and research priorities.

Aside from the changing course of research over the decades, a second striking characteristic is the vast diversity of the work. At its height, this diversity arguably rivaled that found in the psychology and education departments of major research universities anywhere in the world. Moreover, in some areas—particularly in psychometrics and statistics—it was often considerably deeper.

This breadth and depth led to substantial innovation, as this chapter has highlighted and the prior ones have detailed. That innovation was often highly theoretical—as in Witkin and Goodenough's (1981) work on cognitive styles, Sigel's (1990) distancing theory, Lord and Novick's (1968) seminal volume on IRT, Messick's (1989) unified conception of validity, Mislevy's (1994, 1996) early work on ECD, Deane et al.'s (2015) English language arts competency model, and Sykes and Wilson's (2015) conceptions of teaching practice. But that innovation was also very often practical—witness the in-basket test (Frederiksen et al. 1957), LISREL (Jöreskog and van Thillo 1972), the EM algorithm (Dempster et al. 1977), Lord's (1980) "Applications of Item Response Theory to Practical Testing Problems," the application of Mantel–Haenszel to DIF (Holland and Thayer 1988), the plausible-values solution to the estimation of population performance in sample surveys (Mislevy et al. 1992a), and e-rater (Burstein et al. 1998). These innovations were not only useful but *used*, in all the preceding cases widely employed in the measurement community, and in some cases used throughout the sciences.

Of no small consequence is that ETS innovations—theory and practical development—were employed throughout the organization's history to support, challenge, and improve the technical quality of its testing programs. Among other things, the challenges took the form of a continuing program of validity research to identify and address construct-irrelevant influences, for example, test anxiety, coaching, stereotype threat, lack of computer familiarity, English language complexity in content assessments, and accessibility—which might unfairly affect the performance of individuals and groups.

A final observation is that research was used not only for the generation of theory and of practical solutions in educational and psychological studies but also for helping government officials and the public address important policy problems. The organization's long history of contributions to informing policy are evident in its roles with respect to the Equality of Educational Opportunity Study (Beaton 1968); the evaluation of *Sesame Street* (Ball and Bogatz 1970); the Head Start, early childhood, and high school longitudinal studies; the adult literacy studies; NAEP, PISA, and PIAAC; and the many policy analyses of equity and opportunity in the United States (Kirsch et al. 2007; Kirsch and Braun 2016).

We close this chapter, and the book, by returning to the concept of a nonprofit measurement organization as outlined by Bennett (Chap. 1, this volume). In that conception, the organization's raison d'être is public service. Research plays a fundamental role in realizing that public service obligation to the extent that it helps advance educational and psychological measurement as a field, acts as a mechanism for enhancing (and routinely challenging) the organization's testing programs, and helps contribute to the solution of big educational and social challenges. We would assert that the evidence presented indicates that, taken over its almost 70-year

history, the organization's research activities have succeeded in filling that fundamental role.

References

Ackerman, D. J., & Coley, R. J. (2012). *State pre-K assessment policies: Issues and status* (Policy Information Report). Princeton: Educational Testing Service.

Alderman, D. L. (1978). *Evaluation of the TICCIT computer-assisted instructional system in the community college*. Princeton: Educational Testing Service.

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT verbal scores. *American Educational Research Journal, 17*, 239–251. https://doi.org/10.3102/00028312017002239

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

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140. https://doi.org/10.1007/BF02291180

Anderson, S. B., & Ball, S. (1978). *The profession and practice of program evaluation*. San Francisco: Jossey-Bass.

Anderson, S. B., Beaton, A. E., Emmerich, W., & Messick, S. J. (1968). *Disadvantaged children and their first school experiences: ETS-OEO Longitudinal Study—Theoretical considerations and measurement strategies* (Program Report No. PR-68-04). Princeton: Educational Testing Service.

Anderson, S. B., Ball, S., & Murphy, R. T. (Eds.). (1975). *Encyclopedia of educational evaluation: Concepts and techniques for evaluating education and training programs*. San Francisco: Jossey-Bass.

Angoff, W. H. (1953). *Equating of the ACE psychological examinations for high school students* (Research Bulletin No. RB-53-03). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1953.tb00887.x

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J. M., & Roberts, R. D. (2015). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment, 33*, 83–97. http://dx.doi.org/10.1002/10.1177/0734282914550387

Arieli-Attali, M., & Cayton-Hodges, G. A. (2014). *Expanding the CBAL mathematics assessments to elementary grades: The development of a competency model and a rational number learning progression* (Research Report No. RR-14-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12008

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org/

Baird, L. L. (1973). Teaching styles: An exploratory study of dimensions and effects. *Journal of Educational Psychology, 64*, 15–21. https://doi.org/10.1037/h0034058

Baird, L. L., & Knapp, J. E. (1981). *The inventory of documented accomplishments for graduate admissions: Results of a field trial study of its reliability, short-term correlates, and evaluation* (Research Report No. RR-81-18). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1981.tb01253.x

Baker, A. H., Mishara, B. L., Kostin, I. W., & Parker, L. (1976). Kinesthetic aftereffect and personality: A case study of issues involved in construct validation. *Journal of Personality and Social Psychology, 34*, 1–13. https://doi.org/10.1037/0022-3514.34.1.1

Baker, A. H., Mishara, B. L., Parker, L., & Kostin, I. W. (1978). When "reliability" fails, must a measure be discarded?—The case of kinesthetic aftereffect. *Journal of Research in Personality, 12*, 262–273. https://doi.org/10.1016/0092-6566(78)90053-3

Baker, A. H., Mishara, B. L., Kostin, I. W., & Parker, L. (1979). Menstrual cycle affects kinesthetic aftereffect, an index of personality and perceptual style. *Journal of Personality and Social Psychology, 37*, 234–246. https://doi.org/10.1037/0022-3514.37.2.234

Ball, S., & Bogatz, G. A. (1970). *The first year of* Sesame Street*: An evaluation* (Program Report No. PR-70-15). Princeton: Educational Testing Service.

Ball, S., & Bogatz, G. A. (1973). *Reading with television: An evaluation of* The Electric Company (Program Report No. PR-73-02). Princeton: Educational Testing Service.

Barton, P. E. (2003). *Parsing the achievement gap: Baselines for tracking progress* (Policy Information Report). Princeton: Educational Testing Service.

Barton, P. E., & Coley, R. J. (1990). *The education reform decade* (Policy Information Report). Princeton: Educational Testing Service.

Barton, P. E., & Coley, R. J. (2007). *The family: America's smallest school* (Policy Information Report). Princeton: Educational Testing Service.

Barton, P. E., & Coley, R. J. (2010). *The Black–White achievement gap: When progress stopped* (Policy Information Report). Princeton: Educational Testing Service.

Barton, P. E., Coley, R. J., & Goertz, M. E. (1991). *The state of inequality* (Policy Information Report). Princeton: Educational Testing Service.

Beaton, A. E. (1968). *The computer techniques used in the equality of Educational Opportunity Survey* (Research Memorandum No. RM-68-16). Princeton: Educational Testing Service.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522–532. https://doi.org/10.1037/0021-9010.76.4.522

Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation—Final report to the National Council of Architectural Registration Boards* (Research Memorandum No. RM-99-02). Princeton: Educational Testing Service.

Bejar, I. I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton: Educational Testing Service.

Bell, C. A., Wilson, S. M., Higgins, T., & McCoach, D. B. (2010). Measuring the effects of professional development on teacher knowledge: The case of developing mathematical ideas. *Journal for Research in Mathematics Education, 41*, 479–512.

Bennett, R. E. (1999). Computer-based testing for examinees with disabilities: On the Road to generalized accommodation. In S. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 181–191). Mahwah: Erlbaum.

Bennett, R. E. (2016). *Opt out: An examination of issues* (RR-16-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12101

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer. https://doi.org/10.1007/978-1-4020-9964-9_3

Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered Formulating-Hypotheses test. *Journal of Educational Measurement, 32*, 19–36. https://doi.org/10.1111/j.1745-3984.1995.tb00454.x

Bennett, R. E., & Ward, W. C. (Eds). (1993). *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Erlbaum.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

Bennett, R. E., Deane, P., & van Rijn, P. W. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist, 51*, 82–107. https://doi.org/10.1080/00461520.2016.1141683

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*, 294–309. https://doi.org/10.1177/01466210022031769

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007–466). Washington, DC: National Center for Education Statistics.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*, 163–177. https://doi.org/10.1111/j.1745-3984.1997.tb00512.x

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models (PIC-VAM)*. Princeton, NJ: Educational Testing Service.

Braun, H. I., Bennett, R. E., Frye, D, & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement, 27*, 93–108. https://doi.org/10.1111/j.1745-3984.1990.tb00736.x

Berke, J. S., Goertz, M. E., & Coley, R. J. (1984). *Politicians, judges, and city schools: Reforming school finance in New York*. New York: Russell Sage Foundation.

Bhat, S., & Yoon, S-Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication, 67*, 42–57. https://doi.org/10.1016/j.specom.2014.09.005

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review* (Report No. 83–6). New York: College Entrance Examination Board.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill. New York: College Entrance Examination Board.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement history examination. *Journal of Educational Measurement, 31*(4), 275–293. https://doi.org/10.1111/j.1745-3984.1994.tb00447.x

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37–50. https://doi.org/10.1111/j.1745-3984.1994.tb00433.x

Bridgeman, B., Burton, N., & Pollack, J. (2008). Predicting grades in college courses: A comparison of multiple regression and percent succeeding approaches. *Journal of College Admission, 199*, 19–25.

Brinker, R. P., & Lewis, M. (1982). Discovering the competent handicapped infant: A process approach to assessment and intervention. *Topics in Early Childhood Special Education, 2*(2), 1–16. https://doi.org/10.1177/027112148200200205

Brinker, R. P., & Thorpe, M. E. (1984). *Evaluation of the integration of severely handicapped students in education and community settings* (Research Report No. RR-84-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1984.tb00051.x

Brody, G. H., Pellegrini, A. D., & Sigel, I. E. (1986). Marital quality and mother–child and father–child interactions with school-aged children. *Developmental Psychology, 22*, 291–296. https://doi.org/10.1037/0012-1649.22.3.291

Brooks, J., & Lewis, M. (1976). Infants' responses to strangers: Midget, adult, and child. *Child Development, 47*, 323–332. https://doi.org/10.2307/1128785

Brooks-Gunn, J. (1984). The psychological significance of different pubertal events in young girls. *Journal of Early Adolescence, 4*, 315–327. https://doi.org/10.1177/0272431684044003

Brooks-Gunn, J. (1990). Adolescents as daughters and as mothers: A developmental perspective. In I. E. Sigel & G. H. Brody (Eds.), *Methods of family research: Biographies of research projects* (Vol. 1, pp. 213–248). Hillsdale: Erlbaum.

Brooks-Gunn, J., & Hearn, R. P. (1982). Early intervention and developmental dysfunction: Implications for pediatrics. *Advances in Pediatrics, 29*, 497–527.

Brooks-Gunn, J., & Lewis, M. (1981). Infant social perceptions: Responses to pictures of parents and strangers. *Developmental Psychology, 17*, 647–649. https://doi.org/10.1037/0012-1649.17.5.647

Brooks-Gunn, J., & Lewis, M. (1982). Temperament and affective interaction in handicapped infants. *Journal of Early Intervention, 5*, 31–41. https://doi.org/10.1177/105381518200500105

Brooks-Gunn, J., & Warren, M. P. (1988). The psychological significance of secondary sexual characteristics in nine- to eleven-year-old girls. *Child Development, 59*, 1061–1069. https://doi.org/10.2307/1130272

Brooks-Gunn, J., McCormick, M. C., & Heagarty, M. C. (1988). Preventing infant mortality and morbidity: Developmental perspectives. *American Journal of Orthopsychiatry, 58*, 288–296. https://doi.org/10.1111/j.1939-0025.1988.tb01590.x

Bruschi, B. A., & Coley, R. J. (1999). *How teachers compare: The prose, document, and quantitative literacy of America's teachers* (Policy Information Report). Princeton: Educational Testing Service.

Bryan, J. H., & Test, M. A. (1967). Models and helping: Naturalistic studies in aiding behavior. *Journal of Personality and Social Psychology, 6*, 400–407. https://doi.org/10.1037/h0024826

Burrus, J., MacCann, C., Kyllonen, P. C., & Roberts, R. D. (2011). Noncognitive constructs in K-16: Assessments, interventions, educational and policy implications. In P. J. Bowman & E. P. S. John (Eds.), *Readings on equal education: Diversity, merit, and higher education—Toward a comprehensive agenda for the twenty-first century* (Vol. 25, pp. 233–274). New York: AMS Press.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O*NET)* (Research Report No. RR-13-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02328.x

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B. A., Kukich, K.,. .. Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT Analytical Writing Assessment essays* (Research Report No. RR-98-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333–8504.1998.tb01764.x

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (Research Memorandum No. RM-00-06). Princeton: Educational Testing Service.

Buzick, H. M., & Jones, N. D. (2015). Using test scores from students with disabilities in teacher evaluation. *Educational Measurement: Issues and Practice, 34*(3), 28–38. https://doi.org/10.1111/emip.12076

Buzick, H. M., & Laitusis, C. (2010a). *A summary of models and standards-based applications for grade-to-grade growth on statewide assessments and implications for students with disabilities* (Research Report No. RR-10-14). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02221.x

Buzick, H. M., & Laitusis, C. C. (2010b). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*, 537–544. https://doi.org/10.3102/0013189X10383560

Buzick, H. M., & Stone, E. (2011). *Recommendations for conducting differential item functioning (DIF) analyses for students with disabilities based on previous DIF studies* (Research Report No. RR-11-34). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02270.x

Buzick, H. M., Oliveri, M. E., Attali, Y., & Flor, M. (2016). Comparing human and automated essay scoring for prospective graduate students with learning disabilities and/or ADHD. *Applied Measurement in Education, 29*, 161–172. http://dx.doi.org/10.1080/08957347.2016.1171765

Camp, R. (1985). The writing folder in post-secondary assessment. In P. J. A. Evans (Ed.), *Directions and misdirections in English evaluation* (pp. 91–99). Ottawa: Canadian Council of Teachers of English.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 183–212). Hillsdale: Erlbaum.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Campbell, J. T., Esser, B. F., & Flaugher, R. L. (1982). *Evaluation of a program for training dentists in the care of handicapped patients* (Research Report No. RR-82-52). Princeton: Educational Testing Service.

Carnevale, A. P. (1996). Liberal education and the new economy. *Liberal Education, 82*(2), 1–8.

Carnevale, A. P., & Desrochers, D. M. (1997). The role of community colleges in the new economy. *Community College Journal, 67*(5), 25–33.

Carnevale, A. P., & Fry, R. A. (2001). Economics, demography and the future of higher education policy. In *Higher expectations: Essays on the future of postsecondary education* (pp. 13–26). Washington, DC: National Governors Association.

Carnevale, A. P., & Fry, R. A. (2002). The demographic window of opportunity: College access and diversity in the new century. In D. E. Heller (Ed.), *Condition of access: Higher education for lower income students* (pp. 137–151). Westport: Praeger.

Carnevale, A. P., & Rose, S. J. (2000). Inequality and the new high-skilled service economy. In J. Madrick (Ed.), *Unconventional wisdom: Alternative perspectives on the new economy* (pp. 133–156). New York: Century Foundation Press.

Casabianca, J. M., & Junker, B. (2016). Multivariate normal distribution. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 35–46). Boca Raton: CRC Press. https://doi.org/10.1177/0013164413486987

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*, 757–783.

Centra, J. A. (1974). The relationship between student and alumni ratings of teachers. *Educational and Psychological Measurement, 34*, 321–325. https://doi.org/10.1177/001316447403400212

Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education, 18*, 379–389. https://doi.org/10.1007/BF00974804

Centra, J. A., & Barrows, T. S. (1982). *An evaluation of the University of Oklahoma advanced programs: Final report* (Research Memorandum No. RM-82-03). Princeton: Educational Testing Service.

Centra, J. A., & Potter, D. A. (1980). School and teacher effects: An interrelational model. *Review of Educational Research, 50*, 273–291. https://doi.org/10.3102/00346543050002273

Chen, L., Feng, G., Joe, J. N., Leong, C. W., Kitchen, C., & Lee, C. M. (2014a). Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 200–203). New York: ACM. http://dx.doi.org/10.1145/2663204.2663265

Chen, H., Yamamoto, K., & von Davier, M. (2014b). Controlling multistage testing exposure rates in international large-scale assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 391–409). Boca Raton: CRC Press.

Chen, L., Yoon, S.-Y., Leong, C. W., Martin, M., & Ma, M. (2014c). An initial analysis of structured video interviews by using multimodal emotion detection. In *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems* (pp. 1–6). New York: Association of Computational Linguistics. https://doi.org/10.1145/2668056.2668057

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Clewell, B. C. (1987). *Retention of Black and Hispanic doctoral students* (GRE Board Research Report No. 83-4R). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00214.x

Clewell, B. C., Anderson, B. T., & Thorpe, M. E. (1992). *Breaking the barriers: Helping female and minority students succeed in mathematics and science*. San Francisco: Jossey-Bass.

Coley, R. J. (2001). *Differences in the gender gap: Comparisons across racial/ethnic groups in education and work* (Policy Information Report). Princeton: Educational Testing Service.

Coley, R. J., Cradler, J., & Engel, P. (1997). *Computers and classrooms: The status of technology in U.S. schools* (Policy Information Report). Princeton: Educational Testing Service.

Cook, L. L., Eignor, D. R., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using factor analysis to investigate accommodations used by students with disabilities on an English-language arts assessment. *Applied Measurement in Education, 23*, 187–208. https://doi.org/10.1080/08957341003673831

Cumming, A., Kantor, R., Powers, D. E., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series No. TOEFL-MS-18). Princeton: Educational Testing Service.

Damarin, F., & Messick, S. (1965). *Response styles and personality variables: A theoretical integration of multivariate research* (Research Bulletin No. RB-65-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1965.tb00967.x

Danielson, C. (2013). *Framework for teaching evaluation instrument*. Princeton: The Danielson Group.

Danielson, C., & Dwyer, C. A. (1995). How PRAXIS III® supports beginning teachers. *Educational Leadership, 52*(6), 66–67.

Deane, P. (2013a). Covering the construct: An approach to automated essay scoring motivated by socio-cognitive framework for defining literacy skills. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current application and new directions* (pp. 298–312). New York: Routledge.

Deane, P. (2013b). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Deane, P., & Song, Y. (2015). *The key practice, discuss and debate ideas: Conceptual framework, literature review, and provisional learning progressions for argumentation* (Research Report No. RR-15-33). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12079

Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (Research Report No. RR-15-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12071

Deane, P., Sabatini, J. P., Feng, G., Sparks, J. R., Song, Y., Fowles, M. E.,… Foley, C. (2015). *Key practices in the English language arts (ELA): Linking learning theory, assessment, and instruction* (Research Report No. RR-15-17). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12063

Dear, R. E. (1958). *The effects of a program of intensive coaching on SAT scores* (Research Bulletin No. RR-58-05). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1958.tb00080.x

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational and Behavioral Statistics, 11*, 183–196. https://doi.org/10.3102/10769986011003183

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol 26. Psychometrics* (pp. 979–1030). Amsterdam: Elsevier. https://doi.org/10.1016/s0169-7161(06)26031-0

Diederich, P. B. (1957). *The improvement of essay examinations* (Research Memorandum No. RM-57-03). Princeton: Educational Testing Service.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin No. RB-61-15). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1961.tb00286.x

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309–319. https://doi.org/10.1111/j.1745-3984.1992.tb00379.x

Ekstrom, R. B., & Smith, D. K. (Eds.). (2002). *Assessing individuals with disabilities in educational, employment, and counseling settings*. Washington, DC: American Psychological Association. https://doi.org/10.1037/10471-000

Ekstrom, R. B., French, J. W., & Harman, H. H. (with Dermen, D.). (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton: Educational Testing Service.

Ekstrom, R. B., Goertz, M. E., Pollack, J. C., & Rock, D. A. (1991). *Undergraduate debt and participation in graduate education: The relationship between educational debt and graduate school aspirations, applications, and attendance among students with a pattern of full-time continuous postsecondary education* (GRE Research Report No. 86-5). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01330.x

Elliot, N. (2014). *Henry Chauncey: An American life*. New York: Lang.

Emmerich, W. (1973). *Disadvantaged children and their first school experiences: ETS–Head Start longitudinal study—Preschool personal–social behaviors: Relationships with socioeconomic status, cognitive skills, and tempo* (Program Report No. PR-73-33). Princeton: Educational Testing Service.

Enright, M. K., Grabe, W., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series No. MS-17). Princeton: Educational Testing Service.

Enright, M. K., Rock, D. A., & Bennett, R. E. (1998). Improving measurement for graduate admissions. *Journal of Educational Measurement, 35*, 250–267. https://doi.org/10.1111/j.1745-3984.1998.tb00538.x

Eriksson, K., & Haggstrom, O. (2014). Lord's paradox in a continuous setting and a regression artifact in numerical cognition research. *PLoS ONE, 9*(4). https://doi.org/10.1371/journal.pone.0095949

Evans, F. R., & Frederiksen, N. (1974). Effects of models of creative performance on ability to formulate hypotheses. *Journal of Educational Psychology, 66*, 67–82. https://doi.org/10.1037/h0035808

Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10*, 257–272. https://doi.org/10.1111/j.1745-3984.1973.tb00803.x

Feinman, S., & Lewis, M. (1983). Social referencing at ten months: A second-order effect on infants' responses to strangers. *Child Development, 50*, 848–853. https://doi.org/10.2307/1129892

Fife, J. H. (2013). *Automated scoring of mathematics tasks in the Common Core Era: Enhancements to m-rater™ in support of CBAL mathematics and the Common Core Assessments* (Research Report No. RR-13-26). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02333.x

Follman, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53*, 553–562. https://doi.org/10.1007/BF02294407

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*, 476–486. https://doi.org/10.1080/01621459.1992.10475229

Frederiksen, N. O. (1948). *The prediction of first term grades at Hamilton College* (Research Bulletin No. RB-48-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1948.tb00867.x

Frederiksen, N. (1959). *Development of the test "Formulating Hypotheses": A progress report* (Office of Naval Research Technical Report No. NR-2338[00]). Princeton: Educational Testing Service.

Frederiksen, N. O. (1962). Factors in in-basket performance. *Psychological Monographs: General and Applied, 76*(22), 1–25. https://doi.org/10.1037/h0093838

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202. https://doi.org/10.1037/0003-066X.39.3.193

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity and scientific problem-solving. *Applied Psychological Measurement, 2*, 1–24. https://doi.org/10.1177/014662167800200101

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

Frederiksen, N., Saunders, D. R., & Wand, B. (1957). The in-basket test. *Psychological Monographs: General and Applied, 71*(9), 1–28. https://doi.org/10.1037/h0093706

Freedle, R., & Lewis, M. (1977). Prelinguistic conversations. In M. Lewis & L. A. Rosenblum (Eds.), *The origins of behavior: Vol. 5. Interaction, conversation, and the development of language* (pp. 157–185). New York: Wiley.

French, J. W. (1948). The validity of a persistence test. *Psychometrika, 13*, 271–277. https://doi.org/10.1007/BF02289223

French, J. W. (1954). *Manual for Kit of Selected Tests for Reference Aptitude and Achievement Factors*. Princeton: Educational Testing Service.

French, J. W. (1956). *The effect of essay tests on student motivation* (Research Bulletin No. RB-56-04). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1956.tb00060.x

French, J. W. (1958). Validation of new item types against four-year academic criteria. *Journal of Educational Psychology, 49*, 67–76. https://doi.org/10.1037/h0046064

French, J. W. (1962). Effect of anxiety on verbal and mathematical examination scores. *Educational and Psychological Measurement, 22*, 555–567. https://doi.org/10.1177/001316446202200313

French, J. W., & Dear, R. E. (1959). Effect of coaching on an aptitude test. *Educational and Psychological Measurement, 19*, 319–330. https://doi.org/10.1177/001316445901900304

French, J. W., Tucker, L. R., Newman, S. H., & Bobbitt, J. M. (1952). A factor analysis of aptitude and achievement entrance tests and course grades at the United States Coast Guard Academy. *Journal of Educational Psychology, 43*, 65–80. https://doi.org/10.1037/h0054549

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual for Kit of Reference Tests for Cognitive Factors*. Princeton: Educational Testing Service.

Furstenberg, F. F., Jr., Brooks-Gunn, J., & Morgan, S. P. (1987). *Adolescent mothers in later life*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511752810

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended computerized mathematics task. *Educational Assessment, 8*, 27–41. https://doi.org/10.1207/S15326977EA0801_02

Gardner, R. W., Jackson, D. N., & Messick, S. (1960). Personality organization in cognitive controls and intellectual abilities. *Psychological Issues, 2*(4, Whole No. 8). https://doi.org/10.1037/11215-000

Gitomer, D. H. (2007a). *The impact of the National Board for Professional Teaching Standards: A review of the research* (Research Report No. RR-07-33). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02075.x

Gitomer, D. H. (2007b). *Teacher quality in a changing policy landscape: Improvements in the teacher pool* (Policy Information Report). Princeton: Educational Testing Service.

Gitomer, D. H. (2009). *Measurement issues and assessment for teaching quality*. Los Angeles: Sage. https://doi.org/10.4135/9781483329857

Gitomer, D. H., & Bell, C. A. (2013). Evaluating teachers and teaching. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 415–444). Washington, DC: American Psychological Association. https://doi.org/10.1037/14049-020

Gitomer, D. H., & Bell, C. A. (2016). *Handbook of research on teaching* (5th ed.). Washington, DC: American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6

Gitomer, D. H., & Yamamoto, K. (1991). *Performance modeling that integrates latent trait and class theory* (Research Report No. RR-91-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01367.x

Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1994). *Diagnostic assessment of troubleshooting skill in an intelligent tutoring system* (Research Report No. RR-94-21-ONR). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1994.tb01594.x

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis* (NCCTQ Report). Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L. (2013). Can teacher evaluation improve teaching? *Principal Leadership, 13*(7), 24–29.

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness* (NCCTQ Research-to-Practice Brief). Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis* (NCCTQ Report). Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning* (NCCTQ Research & Policy Brief). Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Holdheide, L., & Miller, T. (2011). *A practical guide to designing comprehensive teacher evaluation systems: A tool to assist in the development of teacher evaluation systems* (NCCTQ Report). Washington, DC: National Comprehensive Center for Teacher Quality.

Goertz, M. E. (1978). *Money and education: Where did the 400 million dollars go? The impact of the New Jersey Public School Education Act of 1975*. Princeton: Educational Testing Service.

Goertz, M. E. (1989). *What Americans study* (Policy Information Report). Princeton: Educational Testing Service.

Goertz, M. E., & Moskowitz, J. (1978). *Plain talk about school finance*. Washington, DC: National Institute of Education.

Goertz, M. E., Ekstrom, R., & Coley, R. (1984). *The impact of state policy on entrance into the teaching profession*. Princeton: Educational Testing Service.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York: Bantam.

Goodenough, D. R., Oltman, P. K., & Cox, P. W. (1987). The nature of individual differences in field dependence. *Journal of Research in Personality, 21*, 81–99. https://doi.org/10.1016/0092-6566(87)90028-6

Graf, E. A. (2009). *Defining mathematics competency in the service of cognitively based assessment for Grades 6 through 8* (Research Report No. RR-09-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02199.x

Green, B. F., Jr. (1950a). *A general solution for the latent class model of latent structure analysis* (Research Bulletin No. RB-50-38). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00917.x

Green, B. F., Jr. (1950b). *Latent structure analysis and its relation to factor analysis* (Research Bulletin No. RB-50-65). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00920.x

Gu, L., Lockwood, J., & Powers, D. E. (2015). *Evaluating the TOEFL Junior® Standard Test as a measure of progress for young English language learners* (Research Report No. RR-15-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12064

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley. https://doi.org/10.1037/13240-000

Guzman-Orth, D., Laitusis, C., Thurlow, M., & Christensen, L. (2016). *Conceptualizing accessibility for English language proficiency assessments* (Research Report No. RR-16-07). Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12093

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5*, 815–841. https://doi.org/10.1214/aos/1176343941

Haberman, S. J. (1988). A stabilized Newton–Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology, 18*, 193–211. https://doi.org/10.2307/271049

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. (2010). *Limits on the accuracy of linking* (Research Report No. RR-10-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02229.x

Haberman, S. J. (2016). Exponential family distributions relevant to IRT. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 47–69). Boca Raton: CRC Press.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 209–227. https://doi.org/10.1007/s11336-010-9158-4

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report No. RR-08-45). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02131.x

Handwerk, P., Tognatta, N., Coley, R. J., & Gitomer, D. H. (2008). *Access to success: Patterns of Advanced Placement participation in U.S. high schools* (Policy Information Report). Princeton: Educational Testing Service.

Hansen, E. G., & Mislevy, R. J. (2006). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. L. Howell (Eds.), *Online and distance learning: Concepts, methodologies, tools, and applications* (pp. 214–261). Hershey: Information Science.

Hansen, E. G., Forer, D. C., & Lee, M. J. (2004). *Toward accessible computer-based tests: Prototypes for visual and other disabilities* (Research Report No. RR-04-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01952.x

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System, 33*(1), 107–133. https://doi.org/10.1016/j.system.2004.11.002

Hansen, E. G., Laitusis, C. C., Frankel, L., & King, T. C. (2012). Designing accessible technology-enabled reading assessments: Recommendations from teachers of students with visual impairments. *Journal of Blindness Innovation and Research, 2*(2). http://dx.doi.org/10.5241/2F2-22

Hauck, M. C., Wolf, M. K., & Mislevy, R. (2016). *Creating a next-generation system of K–12 English learner language proficiency assessments* (Research Report No. RR-16-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12092

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (pp. 173–190). Madison: Springer International. https://doi.org/10.1007/978-3-319-19977-1_13

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (Vol. 2, pp. 749–776). Hershey: Information Science Reference. http://dx.doi.org/10.4018/978-1-4666-9441-5.ch029

Heilman, M., & Madnani, N. (2012). Discriminative edit models for paraphrase scoring. In *Proceedings of the first Joint Conference on Lexical and Computational Semantics* (pp. 529–535). Stroudsburg: Association of Computational Linguistics.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks: Sage.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210. https://doi.org/10.1007/s11336-008-9089-5

Hills, J. R. (1958). Needs for achievement, aspirations, and college criteria. *Journal of Educational Psychology, 49*, 156–161. https://doi.org/10.1037/h0047283

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–970. https://doi.org/10.1080/01621459.1986.10478354

Holland, P. W. (1987). *Which comes first, cause or effect?* (Research Report No. RR-87-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00212.x

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 3–25). Hillsdale: Erlbaum.

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

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (Program Statistics Research Technical Report No. 89-84). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00333.x

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale: Erlbaum.

Holland, P. W., King, B. F., & Thayer, D. T. (1989). *The standard error of equating for the kernel method of equating score distributions* (Research Report No. RR-89-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1989.tb00332.x

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2) Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/

Huddleston, E. M. (1952). *Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques* (Research Bulletin No. RB-52-07). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1952.tb00925.x

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*, 243–252. https://doi.org/10.1037/h0045996

Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement, 21*, 771–790. https://doi.org/10.1177/001316446102100402

Johnson, M. S., & Jenkins, F. (2005). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress* (Research Report No. RR-04-38). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01965.x

Jöreskog, K. G. (1965). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31*, 165–178. https://doi.org/10.1007/BF02289505

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239–251. https://doi.org/10.1093/biomet/57.2.239

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426. https://doi.org/10.1007/BF02291366

Jöreskog, K. G., & van Thillo, M. (1972). *LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (Research Bulletin No. RB-72-56). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1972.tb00827.x

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement, 48*, 12–30. https://doi.org/10.1111/j.1745-3984.2010.00128.x

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*, 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64–80). New York: Routledge.

Kaplan, R. M. (1992). *Using a trainable pattern-directed computer program to score natural language item responses* (Research Report No. RR-91-31). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01398.x

Kaplan, R. M., Burstein, J., Trenholm, H., Lu, C., Rock, D., Kaplan, B., & Wolff, C. (1995). *Evaluating a prototype essay scoring procedure using off-the shelf software* (Research Report No. RR-95-21). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01656.x

Karon, B. P., & Cliff, R. H., & (1957). *The Cureton–Tukey method of equating test scores* (Research Bulletin No. RB-57-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1957.tb00072.x

Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers: World Book.
Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49,* 2, p. 161–177. http://dx.doi.org/10.1080/00273171.2013.866536
Kirsch, I., & Braun, H. (Eds.). (2016). *The dynamics of opportunity in America.* New York: Springer. https://doi.org/10.1007/978-3-319-25991-8
Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults* (NAEP Report No. 16-PL-01). Princeton: National Assessment of Educational Progress.
Kirsch, I. S., Braun, H., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future* (Policy Information Report). Princeton: Educational Testing Service.
Kirsch, I., Braun, H., Lennon, M. L., & Sands, A. (2016). *Choosing our future: A story of opportunity in America*. Princeton: Educational Testing Service.
Klein, S. P., Frederiksen, N., & Evans, F. R. (1969). Anxiety and learning to formulate hypotheses. *Journal of Educational Psychology, 60*, 465–475. https://doi.org/10.1037/h0028351
Kogan, N., & Doise, W. (1969). Effects of anticipated delegate status on level of risk taking in small decision-making groups. *Acta Psychologica, 29*, 228–243. https://doi.org/10.1016/0001-6918(69)90017-1
Kogan, N., & Morgan, F. T. (1969). Task and motivational influences on the assessment of creative and intellective ability in children. *Genetic Psychology Monographs, 80*, 91–127.
Kogan, N., & Pankove, E. (1972). Creative ability over a five-year span. *Child Development, 43*, 427–442. https://doi.org/10.2307/1127546
Kogan, N., & Wallach, M. A. (1964). *Risk taking: A study in cognition and personality*. New York: Holt, Rinehart, and Winston.
Kogan, N., Lamm, H., & Trommsdorff, G. (1972). Negotiation constraints in the risk-taking domain: Effects of being observed by partners of higher or lower status. *Journal of Personality and Social Psychology, 23*, 143–156. https://doi.org/10.1037/h0033035
Laitusis, C. C., Mandinach, E. B., & Camara, W. J. (2002). *Predictive validity of SAT I Reasoning Test for test-takers with learning disabilities and extended time accommodations* (Research Report No. RR-02-11). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2002.tb01878.x
Lamm, H., & Kogan, N. (1970). Risk taking in the context of intergroup negotiation. *Journal of Experimental Social Psychology, 6*, 351–363. https://doi.org/10.1016/0022-1031(70)90069-7
Laosa, L. M. (1978). Maternal teaching strategies in Chicano families of varied educational and socioeconomic levels. *Child Development, 49*, 1129–1135. https://doi.org/10.2307/1128752
Laosa, L. M. (1980a). Maternal teaching strategies and cognitive styles in Chicano families. *Journal of Educational Psychology, 72*, 45–54. https://doi.org/10.1037/0022-0663.72.1.45
Laosa, L. M. (1980b). Maternal teaching strategies in Chicano and Anglo-American families: The influence of culture and education on maternal behavior. *Child Development, 51*, 759–765. https://doi.org/10.2307/1129462
Laosa, L. M. (1984). Ethnic, socioeconomic, and home language influences upon early performance on measures of abilities. *Journal of Educational Psychology, 76*, 1178–1198. https://doi.org/10.1037/0022-0663.76.6.1178
Laosa, L. M. (1990). Psychosocial stress, coping, and development of Hispanic immigrant children. In F. C. Serafica, A. I. Schwebel, R. K. Russell, P. D. Isaac, & L. B. Myers (Eds.), *Mental health of ethnic minorities* (pp. 38–65). New York: Praeger.
Leacock, C., & Chodorow, M. (2003). *c-rater*: Scoring of short-answer questions. *Computers and the Humanities, 37*, 389–405. https://doi.org/10.1023/A:1025779619903
Lee, V. E., Brooks-Gunn, J., & Schnur, E. (1988). Does Head Start work? A 1-year follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Developmental Psychology, 24*, 210–222. https://doi.org/10.1037/0012-1649.24.2.210
LeMahieu, P. G., Gitomer, D. H., & Eresh, J. A. T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11–28. https://doi.org/10.1111/j.1745-3992.1995.tb00863.x

Lennon, M. L., Kirsch, I. S., von Davier, M., Wagner, M., & Yamamoto, K. (2003). *Feasibility study for the PISA ICT Literacy Assessment: Report to Network A* (ICT Literacy Assessment Report). Princeton: Educational Testing Service.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin No. RB-55-23). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1955.tb00266.x

Lewis, M. (1977). Early socioemotional development and its relevance to curriculum. *Merrill-Palmer Quarterly, 23*, 279–286.

Lewis, M. (1978). Attention and verbal labeling behavior in preschool children: A study in the measurement of internal representations. *Journal of Genetic Psychology, 133*, 191–202. https://doi.org/10.1080/00221325.1978.10533377

Lewis, M., & Brooks-Gunn, J. (1979). *Social cognition and the acquisition of self*. New York: Plenum Press. https://doi.org/10.1007/978-1-4684-3566-5

Lewis, M., & Brooks-Gunn, J. (1981a). Attention and intelligence. *Intelligence, 5*, 231–238. https://doi.org/10.1016/S0160-2896(81)80010-4

Lewis, M., & Brooks-Gunn, J. (1981b). Visual attention at three months as a predictor of cognitive functioning at two years of age. *Intelligence, 5*, 131–140. https://doi.org/10.1016/0160-2896(81)90003-9

Lewis, M., & Feiring, C. (1982). Some American families at dinner. In L. Laosa & I. Sigel (Eds.), *Families as learning environments for children* (pp. 115–145). New York: Plenum Press. https://doi.org/10.1007/978-1-4684-4172-7_4

Lewis, M., & Michalson, L. (1982). The measurement of emotional state. In C. E. Izard & P. B. Read (Eds.), *Measuring emotions in infants and children* (Vol. 1, pp. 178–207). New York: Cambridge University Press.

Lewis, M., & Rosenblum, L. A. (Eds.). (1978). *Genesis of behavior: Vol. 1. The development of affect*. New York: Plenum Press.

Lewis, M., Brooks, J., & Haviland, J. (1978). Hearts and faces: A study in the measurement of emotion. In M. Lewis & L. A. Rosenblum (Eds.), *Genesis of behavior: Vol. 1. The development of affect* (pp. 77–123). New York: Plenum Press. https://doi.org/10.1007/978-1-4684-2616-8_4

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics, 34*, 433–463. https://doi.org/10.3102/1076998609332757

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161. https://doi.org/10.3102/00346543043002139

Linn, R. L. (1976). In search of fair selection procedures. *Journal of Educational Measurement, 13*, 53–58. https://doi.org/10.1111/j.1745-3984.1971.tb00898.x

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1–4. https://doi.org/10.1111/j.1745-3984.1971.tb00898.x

Lipnevich, A. A., & Roberts, R. D. (2012). Noncognitive skills in education: Emerging research and applications in a variety of international contexts. *Learning and Individual Differences, 22*, 173–177. https://doi.org/10.1016/j.lindif.2011.11.016

Liu, O. L., Lee, H. S., & Linn, M. C. (2010). An investigation of teacher impact on student inquiry science performance using a hierarchical linear model. *Journal of Research in Science Teaching, 47*, 807–819. https://doi.org/10.1002/tea.20372

Liu, L., Rogat, A., & Bertling, M. (2013). *A CBAL science model of cognition: Developing a competency model and learning progressions to support assessment development* (Research Report No. RR-13-29). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02336.x

Lockheed, M. E. (1985). Women, girls and computers: A first look at the evidence. *Sex Roles, 13*, 115–122. https://doi.org/10.1007/BF00287904

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *Annals of Applied Statistics, 9*, 1484–1509. https://doi.org/10.1214/15-AOAS833

Lopez, A. A., Pooler, E., & Linquanti, R. (2016). *Key issues and opportunities in the initial identification and classification of English learners* (Research Report No. RR-16-09). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12090

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph, 17*(7).

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*, 57–76. https://doi.org/10.1007/BF02289028

Lord, F. M. (1965a). An empirical study of item-test regression. *Psychometrika, 30*, 373–376. https://doi.org/10.1007/BF02289501

Lord, F. M. (1965b). A note on the normal ogive or logistic curve in item analysis. *Psychometrika, 30*, 371–372. https://doi.org/10.1007/BF02289500

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–335. https://doi.org/10.1037/h0025105

Lord, F. M. (1968a). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989–1020. https://doi.org/10.1177/001316446802800401

Lord, F. M. (1968b). *Some test theory for tailored testing* (Research Bulletin No. RB-68-38). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1968.tb00562.x

Lord, F. M. (1970). Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika, 35*, 43–50. https://doi.org/10.1007/BF02290592

Lord, F. M. (1973). Power scores estimated by item characteristic curves. *Educational and Psychological Measurement, 33*, 219–224. https://doi.org/10.1177/001316447303300201

Lord, F. M. (1974a). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247–264. https://doi.org/10.1007/BF02291471

Lord, F. M. (1974b). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 106–126). San Francisco: Freeman.

Lord, F. M. (1975a). The "ability" scale in item characteristic curve theory. *Psychometrika, 40*, 205–217. https://doi.org/10.1007/BF02291567

Lord, F. M. (1975b). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin No. RB-75-33). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01073.x

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117–138. https://doi.org/10.1111/j.1745-3984.1977.tb00032.x

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

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

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotions, 8*, 540–551. https://doi.org/10.1037/a0012746

MacCann, C., Schulze, R., Matthews, G., Zeidner, M., & Roberts, R. D. (2008). Emotional intelligence as pop science, misled science, and sound science: A review and critical synthesis of perspectives from the field of psychology. In N. C. Karafyllis & G. Ulshofer (Eds.), *Sexualized brains: Scientific modeling of emotional intelligence from a cultural perspective* (pp. 131–148). Cambridge, MA: MIT Press.

MacCann, C., Wang, L., Matthews, G., & Roberts, R. D. (2010). Emotional intelligence and the eye of the beholder: Comparing self- and parent-rated situational judgments in adolescents. *Journal of Research in Personality, 44*, 673–676. https://doi.org/10.1016/j.jrp.2010.08.009

MacCann, C., Fogarty, G. J., Zeidner, M., & Roberts, R. D. (2011). Coping mediates the relationship between emotional intelligence (EI) and academic achievement. *Contemporary Educational Psychology, 36*, 60–70. https://doi.org/10.1016/j.cedpsych.2010.11.002

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Marco, G. L. (1972). *Impact of Michigan 1970–71 Grade 3 Title I reading programs* (Program Report No. PR-72-05). Princeton: Educational Testing Service.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*, 160–179. https://doi.org/10.1080/10627190903422906

Mazzeo, J., & von Davier, M. (2008). (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB, 28*, 23–24.

McCaffrey, D. F. (2013). *Will teacher value-added scores change when accountability tests change?* (Carnegie Knowledge Network Brief No. 8). Retrieved from http://www.carnegieknowledgenetwork.org/briefs/valueadded/accountability-tests/

McCaffrey, D. F., Han, B., & Lockwood, J. R. (2014). Using auxiliary teacher data to improve value-added: An application of small area estimation to middle school mathematics teachers. In R. W. Lissitz & H. Jiao (Eds.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness* (pp. 191–217). Charlotte: Information Age.

McDonald, F. J., & Elias, P. (1976). *Beginning teacher evaluation study, Phase 2: The effects of teaching performance on pupil learning* (Vol. 1, Program Report No. PR-76-06A). Princeton: Educational Testing Service.

McGillicuddy-DeLisi, A. V., Sigel, I. E., & Johnson, J. E. (1979). The family as a system of mutual influences: Parental beliefs, distancing behaviors, and children's representational thinking. In M. Lewis & L. A. Rosenblum (Eds.), *Genesis of behavior: Vol. 2. The child and its family* (pp. 91–106). New York: Plenum Press.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*, 242–247. https://doi.org/10.1080/00220671.1987.10885759

Medley, D. M., & Hill, R. A. (1967). Dimensions of classroom behavior measured by two systems of interaction analysis. *Educational Leadership, 26*, 821–824.

Medley, D. M., Coker, H., Lorentz, J. L., Soar, R. S., & Spaulding, R. L. (1981). Assessing teacher performance from observed competency indicators defined by classroom teachers. *Journal of Educational Research, 74*, 197–216. https://doi.org/10.1080/00220671.1981.10885311

Melville, S. D., & Frederiksen, N. (1952). Achievement of freshmen engineering students and the Strong Vocational Interest Blank. *Journal of Applied Psychology, 36*, 169–173. https://doi.org/10.1037/h0059101

Messick, S. (1967). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 115–145). Chicago: Aldine.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1980). *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC*. Princeton: Educational Testing Service.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational testing and practice. *Educational Psychologist, 17*, 67–91. https://doi.org/10.1080/00461528209529246

Messick, S. (1987). Structural relationships across cognition, personality, and style. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction: Vol. 3. Conative and affective process analysis* (pp. 35–75). Hillsdale: Erlbaum.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Messick, S. (1994b). The matter of style: Manifestations of personality in cognition, learning, and teaching. *Educational Psychologist, 29*, 121–136. https://doi.org/10.1207/s15326985ep2903_2

Messick, S. (1996). Bridging cognition and personality in education: The role of style in performance and development. *European Journal of Personality, 10*, 353–376. https://doi.org/10.1002/(SICI)1099-0984(199612)10:5<353::AID-PER268>3.0.CO;2-G

Messick, S., Beaton, A. E., & Lord, F. (1983). *National Assessment of Educational Progress: A new design for a new era*. Princeton: Educational Testing Service.

Messick, S., & Fritzky, F. J. (1963). Dimensions of analytic attitude in cognition and personality. *Journal of Personality, 31*, 346–370. https://doi.org/10.1111/j.1467-6494.1963.tb01304.x

Messick, S., & Kogan, N. (1966). Personality consistencies in judgment: Dimensions of role constructs. *Multivariate Behavioral Research, 1*, 165–175. https://doi.org/10.1207/s15327906mbr0102_3

Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 9–49). San Francisco: Jossey-Bass.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381. https://doi.org/10.1007/BF02306026

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993–997. https://doi.org/10.1080/01621459.1985.10478215

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J. (1993a). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika, 58*, 79–85. https://doi.org/10.1007/BF02294472

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

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483. https://doi.org/10.1177/0013164493053002002

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416. https://doi.org/10.1111/j.1745-3984.1996.tb00498.x

Mislevy, R. J., & Levy, R. (2007). Bayesian psychometric modeling from an evidence centered design perspective. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 839–865). Amsterdam: Elsevier. http://dx.doi.org/10.1016/S0169-7161(06)26026-7

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215. https://doi.org/10.1007/BF02295283

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992a). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R., Johnson, E., & Muraki, E. (1992b). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*, 131−154. https://doi.org/10.2307/1165166

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2000). *Bayes nets in educational assessment: Where do the numbers come from?* (CSE Technical Report No. 518). Los Angeles: UCLA CRESST.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363–389. https://doi.org/10.1207/S15324818AME1504_03

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J., Steinberg, L., Almond, R. G., & Lucas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–82). Mahwah: Erlbaum.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., et al. (2014). *Psychometric considerations in game-based assessment*. Redwood City: GlassLab.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2016). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 23–48). New York: Routledge.

Moses, T. (2016). Loglinear models for observed-score distributions. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 71–85). Boca Raton: CRC Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. https://doi.org/10.1177/014662169201600206

Murphy, R. T. (1973). *Adult functional reading study* (Program Report No. PR-73-48). Princeton: Educational Testing Service.

Murphy, R. T. (1977). *Evaluation of the PLATO 4 computer-based education system: Community college component*. Princeton: Educational Testing Service.

Murphy, R. T. (1988). *Evaluation of* Al Manaahil*: An original Arabic children's television series in reading* (Research Report No. RR-88-45). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00301.x

Murphy, R. T., & Appel, L. R. (1984). *Evaluation of the Writing to Read instructional system, 1982–1984*. Princeton: Educational Testing Service.

Myford, C. M., & Engelhard, G., Jr. (2001). Examining the psychometric quality of the National Board for Professional Teaching Standards Early Childhood/Generalist Assessment System. *Journal of Personnel Evaluation in Education, 15*, 253–285. https://doi.org/10.1023/A:1015453631544

Naemi, B. D., Seybert, J., Robbins, S. B., & Kyllonen, P. C. (2014). *Examining the WorkFORCE® Assessment for Job Fit and Core Capabilities of FACETS* (Research Report No. RR-14-32). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12040

Nettles, M. T. (1990). *Black, Hispanic, and White doctoral students: Before, during, and after enrolling in graduate school* (Minority Graduate Education Project Report No. MGE-90-01). Princeton: Educational Testing Service.

Nettles, M., & Millett, C. (2006). *Three magic letters: Getting to Ph.D*. Baltimore: Johns Hopkins University Press.

Nettles, M. T., Scatton, L. H., Steinberg, J. H., & Tyler, L. L. (2011). *Performance and passing rate differences of African American and White prospective teachers on PRAXIS examinations* (Research Report No. RR-11-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02244.x

Nogee, P. (1950). *A preliminary study of the "Social Situations Test"* (Research Memorandum No. RM-50-22). Princeton: Educational Testing Service.

Oliveri, M. E., & Ezzo, C. (2014). The role of noncognitive measures in higher education admissions. *Journal of the World Universities Forum, 6*(4), 55–65.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315–333.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. http://dx.doi.org/10.1080/15305058.2013.825265

Oranje, A., & Ye, J. (2013). Population model size, bias, and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessments* (pp. 203–228). Boca Raton: CRC Press.

Phelps, G., & Howell, H. (2016). Assessing mathematical knowledge for teaching: The role of teaching context. *The Mathematics Enthusiast, 13*(1), 52–70.

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton: Educational Testing Service.

Powers, D. E. (1985). Effects of test preparation on the validity of a graduate admissions test. *Applied Psychological Measurement, 9*, 179–190. https://doi.org/10.1177/014662168500900206

Powers, D. E. (1988). Incidence, correlates, and possible causes of test anxiety in graduate admissions testing. *Advances in Personality Assessment, 7*, 49–75.

Powers, D. E. (1992). *Assessing the classroom performance of beginning teachers: Educators' appraisal of proposed evaluation criteria* (Research Report No. RR-92-55). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01487.x

Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the Graduate Record Examinations (GRE) General Test. *Journal of Educational Computing Research, 24*, 249–273. https://doi.org/10.2190/680W-66CR-QRP7-CL1F

Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education, 12*, 257–279. https://doi.org/10.1207/S15324818AME1203_3

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76*, 266–278. https://doi.org/10.1037/0022-0663.76.2.266

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*, 407–425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (Research Report No. RR-09-01). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02158.x

Quirk, T. J., Steen, M. T., & Lipe, D. (1971). Development of the Program for Learning in Accordance with Needs Teacher Observation Scale: A teacher observation scale for individualized instruction. *Journal of Educational Psychology, 62*, 188–200. https://doi.org/10.1037/h0031144

Quirk, T. J., Trismen, D. A., Nalin, K. B., & Weinberg, S. F. (1975). The classroom behavior of teachers during compensatory reading instruction. *Journal of Educational Research, 68*, 185–192. https://doi.org/10.1080/00220671.1975.10884742

Quirk, T. J., Witten, B. J., & Weinberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the National Teacher Examinations. *Review of Educational Research, 43*, 89–113. https://doi.org/10.3102/00346543043001089

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (Research Report No. RR-94-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1994.tb01600.x

Rao, C. R., & Sinharay, S. (Eds.). (2006). *Handbook of statistics: Vol. 26. Psychometrics*. Amsterdam: Elsevier.

Renninger, K. A., & Sigel, I. E. (Eds.). (2006). *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed.). New York: Wiley.

Reynolds, A., Rosenfeld, M., & Tannenbaum, R. J. (1992). *Beginning teacher knowledge of general principles of teaching and learning: A national survey* (Research Report No. RR-92-60). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01491.x

Ricciuti, H. N. (1951). *A comparison of leadership ratings made and received by student raters (Research Memorandum No. RM-51-04)*. Princeton: Educational Testing Service.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361–372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 38*, 32–60. https://doi.org/10.3102/1076998614531045

Roberts, R. D., MacCann, C., Matthews, G., & Zeidner, M. (2010). Emotional intelligence: Toward a consensus of models and measures. *Social and Personality Psychology Compass, 4*, 821–840. https://doi.org/10.1111/j.1751-9004.2010.00277.x

Roberts, R. D., Schulze, R., O'Brien, K., McCann, C., Reid, J., & Maul, A. (2006). Exploring the validity of the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) with established emotions measures. *Emotions, 6*, 663–669. https://doi.org/10.1037/1528-3542.6.4.663

Robustelli, S. L. (2010). *Validity evidence to support the development of a licensure assessment for entry-level teachers: A job-analytic approach* (Research Memorandum No. RM-10-10). Princeton: Educational Testing Service.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516–524. https://doi.org/10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics, 41*, 103–116. https://doi.org/10.2307/2530647

Rosenhan, D. (1969). Some origins of concern for others. In P. H. Mussen, J. Langer, & M. V. Covington (Eds.), *Trends and issues in developmental psychology* (pp. 134–153). New York: Holt, Rinehart, and Winston.

Rosenhan, D. (1970). The natural socialization of altruistic autonomy. In J. Macaulay & L. Berkowitz (Eds.), *Altruism and helping behavior: Social psychological studies of some antecedents and consequences* (pp. 251–268). New York: Academic Press.

Rosenhan, D. L. (1972). Learning theory and prosocial behavior. *Journal of Social Issues, 28*, 151–163. https://doi.org/10.1111/j.1540-4560.1972.tb00037.x

Rosenhan, D., & White, G. M. (1967). Observation and rehearsal as determinants of prosocial behavior. *Journal of Personality and Social Psychology, 5*, 424–431. https://doi.org/10.1037/h0024395

Rosner, F. C., & Howey, K. R. (1982). Construct validity in assessing teacher knowledge: New NTE interpretations. *Journal of Teacher Education, 33*(6), 7–12. https://doi.org/10.1177/002248718203300603

Rubin, D. B. (1974a). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association, 69*, 467–474.

Rubin, D. B. (1974b). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1976a). Noniterative least squares estimates, standard errors, and F-tests for analyses of variance with missing data. *Journal of the Royal Statistical Society, Series B, 38*, 270–274.

Rubin, D. B. (1976b). Comparing regressions when some predictor values are missing. *Technometrics, 18*, 201–205. https://doi.org/10.1080/00401706.1976.10489425

Rubin, D. B. (1976c). Inference and missing data. *Biometrika, 63*, 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6*, 34–58. https://doi.org/10.1214/aos/1176344064

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318–328. https://doi.org/10.1080/01621459.1979.10482513

Rubin, D. B. (1980a). Bias reduction using Mahalanobis-metric matching. *Biometrics, 36*, 293–298. https://doi.org/10.2307/2529981

Rubin, D. B. (1980b). *Handling nonresponse in sample surveys by multiple imputations*. Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Rubin, D. B. (1980c). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. *42nd Session of the International Statistical Institute, 1979*(Book 2), 517–532.

Rudd, R., Kirsch, I., & Yamamoto, K. (2004). *Literacy and health in America* (Policy Information Report). Princeton: Educational Testing Service.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. London: Chapman and Hall.

Ryans, D. G., & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 455–494). Washington, DC: American Council on Education.

Sabatini, J., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, L. E. Cutting, & P. McCardle (Eds.), *Unraveling reading comprehension: Behavioral, neurobiological and genetic components* (pp. 100–111). Baltimore: Paul H. Brooks.

Sandoval, J. (1976). *Beginning Teacher Evaluation Study: Phase II. 1973–74. Final report: Vol. 3. The evaluation of teacher behavior through observation of videotape recordings* (Program Report No. PR-76-10). Princeton: Educational Testing Service.

Schrader, W. B., & Pitcher, B. (1964). *Adjusted undergraduate average grades as predictors of law school performance* (Law School Admissions Council Report No. LSAC-64-02). Princeton: LSAC.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Agreement between expert-system and human raters on complex constructed-response quantitative items. *Journal of Applied Psychology, 76*, 856–862. https://doi.org/10.1037/0021-9010.76.6.856

Sherman, S. W., & Robinson, N. M. (Eds.). (1982). *Ability testing of handicapped people: Dilemma for government, science, and the public*. Washington, DC: National Academy Press.

Shipman, V. C. (Ed.). (1972). *Disadvantaged children and their first school experiences: ETS–Head Start longitudinal study* (Program Report No. 72-27). Princeton: Educational Testing Service.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2008). *Monitoring and fostering learning through games and embedded assessments* (Research Report No. RR-08-69). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02155.x

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). New York: Routledge.

Sigel, I. E. (1982). The relationship between parental distancing strategies and the child's cognitive behavior. In L. M. Laosa & I. E. Sigel (Eds.), *Families as learning environments for children* (pp. 47–86). New York: Plenum Press. https://doi.org/10.1007/978-1-4684-4172-7_2

Sigel, I. E. (1990). Journeys in serendipity: The development of the distancing model. In I. E. Sigel & G. H. Brody (Eds.), *Methods of family research: Biographies of research projects: Vol. 1. Normal families* (pp. 87–120). Hillsdale: Erlbaum.

Sigel, I. E. (1992). The belief–behavior connection: A resolvable dilemma? In I. E. Sigel, A. V. McGillicuddy-DeLisi, & J. J. Goodnow (Eds.), *Personal belief systems: The psychological consequences for children* (2nd ed., pp. 433–456). Hillsdale: Erlbaum.

Sigel, I. E. (1993). The centrality of a distancing model for the development of representational competence. In R. R. Cocking & K. A. Renninger (Eds.), *The development and meaning of psychological distance* (pp. 141–158). Hillsdale: Erlbaum.

Sigel, I. E. (1999). Approaches to representation as a psychological construct: A treatise in diversity. In I. E. Sigel (Ed.), *Development of mental representation: Theories and applications* (pp. 3–12). Mahwah: Erlbaum.

Sigel, I. (2000). Educating the Young Thinker model, from research to practice: A case study of program development, or the place of theory and research in the development of educational programs. In J. L. Roopnarine & J. E. Johnson (Eds.), *Approaches to early childhood education* (3rd ed., pp. 315–340). Upper Saddle River: Merrill.

Sigel, I. E. (2006). Research to practice redefined. In K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed., pp. 1017–1023). New York: Wiley.

Sinharay, S. (2003). *Practical applications of posterior predictive model checking for assessing fit of common item response theory models* (Research Report No. RR-03-33). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01925.x

Sinharay, S. (2016). Bayesian model fit and model comparison. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 379–394). Boca Raton: CRC Press.

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BGROUP program to higher dimensions* (Research Report No. RR-05-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb02004.x

Skager, R. W., Schultz, C. B., & Klein, S. P. (1965). Quality and quantity of accomplishments as measures of creativity. *Journal of Educational Psychology, 56*, 31–39. https://doi.org/10.1037/h0021901

Skager, R. W., Schultz, C. B., & Klein, S. P. (1966). Points of view about preference as tools in the analysis of creative products. *Perceptual and Motor Skills, 22*, 83–94. https://doi.org/10.2466/pms.1966.22.1.83

Sparks, J. R., & Deane, P. (2015). *Cognitively based assessment of research and inquiry skills: Defining a key practice in the English language arts* (Research Report No. RR-15-35). Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12082

Sprinthall, N. A., & Beaton, A. E. (1966). Value differences among public high school teachers using a regression model analysis of variance technique. *Journal of Experimental Education, 35*, 36–42. https://doi.org/10.1080/00220973.1966.11010982

Steinberg, L. S., & Gitomer, D. H. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24*, 223–258. https://doi.org/10.1007/BF00119978

Steinberg, J., Cline, F., & Sawaki, Y. (2011). *Examining the factor structure of a state standards-based science assessment for students with learning disabilities* (Research Report No. RR-11-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02274.x

Stone, E., & Davey, T. (2011). *Computer-adaptive testing for students with disabilities: A review of the literature* (Research Report No. RR-11-32). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02268.x

Stone, E., Cook, L. L., & Laitusis, C. (2013). *Evaluation of a condition-adaptive test of reading comprehension for students with reading-based learning disabilities* (Research Report No. RR-13-20). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02327.x

Stone, E., Laitusis, C. C., & Cook, L. L. (2016). Increasing the accessibility of assessments through technology. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 217–234). New York: Routledge.

Stricker, L. J., & Bejar, I. (2004). Test difficulty and stereotype threat on the GRE General Test. *Journal of Applied Social Psychology, 34*, 563–597. https://doi.org/10.1111/j.1559-1816.2004.tb02561.x

Stricker, L. J., & Rock, D. A. (2015). An "Obama effect" on the GRE General Test? *Social Influence, 10*, 11–18. https://doi.org/10.1080/15534510.2013.878665

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665–693. https://doi.org/10.1111/j.1559-1816.2004.tb02561.x

Sum, A., Kirsch, I., & Taggart, R. (2002). *The twin challenges of mediocrity and inequality: Literacy in the U.S. from an international perspective* (Policy Information Report). Princeton: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology, 75*, 104–115. https://doi.org/10.1037/0022-0663.75.1.104

Sykes, G., & Wilson, S. M. (2015). *How teachers teach: Mapping the terrain of practice*. Princeton: Educational Testing Service.

Tannenbaum, R. J. (1992). *A job analysis of the knowledge important for newly licensed (certified) general science teachers* (Research Report No. RR-92-77). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01509.x

Tannenbaum, R. J., & Rosenfeld, M. (1994). Job analysis for teacher competency testing: Identification of basic skills important for all entry-level teachers. *Educational and Psychological Measurement, 54*, 199–211. https://doi.org/10.1177/0013164494054001026

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Thompson, M., & Goe, L. (2009). *Models for effective and scalable teacher professional development* (Research Report No. RR-09-07). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02164.x

Torgerson, W. S., & Green, B. F. (1950). *A factor analysis of English essay readers* (Research Bulletin No. RB-50-30). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1950.tb00470.x

Turkan, S., & Buzick, H. M. (2016). Complexities and issues to consider in the evaluation of content teachers of English language learners. *Urban Education, 51*, 221–248. https://doi.org/10.1177/0042085914543111

U.S. Department of Education. (n.d.-a). *Technical notes on the interactive computer and hands-on tasks in science*. Retrieved from http://www.nationsreportcard.gov/science_2009/ict_tech_notes.aspx

U.S. Department of Education. (n.d.-b). *2011 writing assessment*. Retrieved from http://www.nationsreportcard.gov/writing_2011/

U.S. Department of Education. (n.d.-c). *2014 technology and engineering literacy (TEL) assessment*. Retrieved from http://www.nationsreportcard.gov/tel_2014/

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling and linking*. New York: Springer. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating*. New York: Springer. https://doi.org/10.1007/b97446

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online, 2*(2), 29–48.

von Davier, M. (2008a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

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

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)*, *Research Report No. RR-14-40*. Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12043

von Davier, M. (2016). *High-performance psychometrics: The parallel-E parallel-M algorithm for generalized latent variable models* (Research Report No. 16-34). Princeton: Educational Testing Service.

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 391–406). Boca Raton: CRC Press.

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics, 32*, 233–251. https://doi.org/10.3102/1076998607300422

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam: Elsevier. http://dx.doi.org/10.1016/S0169-7161(06)26032-2

Wagemaker, H., & Kirsch, I. (2008). Editorial. In D. Hastedt & M. von Davier (Eds.), *Issues and methodologies in large scale assessments* IERI monograph series, Vol. 1, pp. 5–7). Hamburg: IERInstitute.

Wainer, H. (Ed.). (1986). *Drawing inferences from self-selected samples.* New York: Springer. https://doi.org/10.1007/978-1-4612-4976-4

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

Walberg, H. J. (1966). *Personality, role conflict, and self-conception in student teachers* (Research Bulletin No. RB-66-10). Princeton: Educational Testing Service.

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity–intelligence distinction*. New York: Holt, Rinehart, and Winston.

Wallach, M. A., Kogan, N., & Bem, D. J. (1962). Group influence on individual risk taking. *Journal of Abnormal and Social Psychology, 65*, 75–86. https://doi.org/10.1037/h0044376

Wang, A. H., Coleman, A. B., Coley, R. J., & Phelps, R. P. (2003). *Preparing teachers around the world* (Policy Information Report). Princeton: Educational Testing Service.

Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of NAACL-HLT 2013* (pp. 814–819). Atlanta: Association of Computational Linguistics.

Ward, L. B. (1960). The business in-basket test: A method of assessing certain administrative skills. *Harvard Business Review, 38*, 164–180.

Ward, W. C. (1973). *Disadvantaged children and their first school experiences: ETS–Head Start Longitudinal Study—Development of self-regulatory behaviors* (Program Report No. PR-73-18). Princeton: Educational Testing Service.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17*, 11–29. https://doi.org/10.1111/j.1745-3984.1980.tb00811.x

Ward, W. C., Kogan, N., & Pankove, E. (1972). Incentive effects in children's creativity. *Child Development, 43*, 669–676. https://doi.org/10.2307/1127565

Wendler, C., Bridgeman, B., Cline, F., Millett, C., Rock, J., Bell, N., & McAllister, P. (2010). *The path forward: The future of graduate education in the United States*. Princeton: Educational Testing Service.

Werts, C. E., Jöreskog, K. G., & Linn, R. L. (1972). A multitrait–multimethod model for studying growth. *Educational and Psychological Measurement, 32*, 655–678. https://doi.org/10.1177/001316447203200308

Werts, C. E., Jöreskog, K. G., & Linn, R. L. (1973). Identification and estimation in path analysis with unmeasured variables. *American Journal of Sociology, 78*, 1469–1484. https://doi.org/10.1086/225474

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah: Erlbaum.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H. I., Rock, D. A., & Powers, D. E. (Eds.). (1988). *Testing handicapped people*. Boston: Allyn and Bacon.

Witkin, H. A., & Goodenough, D. R. (1981). *Cognitive styles: Essence and origins*. New York: International Universities Press.

Witkin, H. A., Moore, C. A., Oltman, P. K., Goodenough, D. R., Friedman, F., Owen, D. R., & Raskin, E. (1977). Role of the field-dependent and field-independent cognitive styles in academic evolution: A longitudinal study. *Journal of Educational Psychology, 69*, 197–211. https://doi.org/10.1037/0022-0663.69.3.197

Witkin, H. A., Price-Williams, D., Bertini, M., Christiansen, B., Oltman, P. K., Ramirez, M., & van Meel, J. M. (1974). Social conformity and psychological differentiation. *International Journal of Psychology, 9*, 11–29. https://doi.org/10.1080/00207597408247089

Wolf, M. K., & Farnsworth, T. (2014). English language proficiency assessments as an exit criterion for English learners. In A. J. Kunnan (Ed.), *The companion to language assessment: Vol. 1. Abilities, contexts, and learners. Part 3, assessment contexts* (pp. 303–317). Wiley-Blackwell: Chichester.

Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English learners: Academic language proficiency and content assessment performance. *Educational Measurement: Issues and Practice, 35*(2), 6–18. https://doi.org/10.1111/emip.12105

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*, 139–159. https://doi.org/10.1080/10627190903425883

Wolf, M. K., Kao, J. C., Rivera, N. M., & Chang, S. M. (2012a). Accommodation practices for English language learners in states' mathematics assessments. *Teachers College Record, 114*(3), 1–26.

Wolf, M. K., Kim, J., & Kao, J. (2012b). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Applied Measurement in Education, 25*, 347–374. https://doi.org/10.1080/08957347.2012.714693

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research* (Research Report No. RR-16-08). Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12091

Wylie, E. C., Lyon, C. J., & Goe, L. (2009). *Teacher professional development focused on formative assessment: Changing teachers, changing schools* (Research Report No. RR-09-10). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02167.x

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*, 147–170. https://doi.org/10.1177/0265532209349465

Xu, X., & von Davier, M. (2008). Linking for the general diagnostic model. In D. Hastedt & M. von Davier (Eds.), *Issues and methodologies in large scale assessments* IERI monograph series, Vol. 1, pp. 97–111). Hamburg: IERInstitute.

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

Yamamoto, K., & Kulick, E. (2002). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 259–277). Chestnut Hill: Boston College.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*, 155–173. https://doi.org/10.2307/1165167

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton: CRC Press.

Young, J. W. (2009). A framework for test validity research on content assessments taken by English language learners. *Educational Assessment, 14*, 122–138. https://doi.org/10.1080/10627190903422856

Young, J. W., & King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies* (Research Report No. RR-08-48). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02134.x

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*, 170–192. https://doi.org/10.1080/10627190802394388

Young, J. W., Steinberg, J., Cline, F., Stone, E., Martiniello, M., Ling, G., & Cho, Y. (2010). Examining the validity of standards-based assessments for initially fluent students and former English language learners. *Educational Assessment, 15*, 87–106. https://doi.org/10.1080/10627197.2010.491070

Young, J. W., King, T. C., Hauck, M. C., Ginsburgh, M., Kotloff, L. J., Cabrera, J., & Cavalie, C. (2014). *Improving content assessment for English language learners: Studies of the linguistic modification of test items* (Research Report No. RR-14-23). Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12023

Zaleska, M., & Kogan, N. (1971). Level of risk selected by individuals and groups when deciding for self and for others. *Sociometry, 34*, 198–213. https://doi.org/10.2307/2786410

Zapata-Rivera, D., & Bauer, M. (2012). Exploring the role of games in educational assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 149–171). Charlotte: Information Age.

Zapata-Rivera, D., Jackson, T., & Katz, I. R. (2014). Authoring conversation-based assessment scenarios. In R. A. Sottilare, A. C. Graesser, X. Hu, & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems* (pp. 169–178). Orlando: U.S. Army Research Laboratory.