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
sense of the progression of topics across time.\footnote{In most cases, citations included as examples of a work stream were selected based on their discussion in one of the book’s chapters.} 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.
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 measure-
ment of personality specifically, although a range of related topics was also investi-
gated, including continued work on response styles (Damarin and Messick 1965;
Jackson and Messick 1961; Messick 1967), the introduction into the social–psycho-
logical 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). 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,

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 Longitudinal 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 Michelson 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 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 sub-score 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).
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.

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,
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., & Messiah, 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/j.1177-0734.2015.00537.x

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., & Burstine, 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
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

Clewel, 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

Clewel, 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

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-Enrt, 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/BF02924407

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/10.1177/014662167800020010

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/10.1177/016028666202200313

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
Advancing Human Assessment: A Synthesis Over Seven Decades

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-Ortiz, 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
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.2330-8504.2004.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–Takey 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 approach 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/10.1016/0001-6918(69)90017-1

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](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](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](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/10.1016/0160-2896(81)90010-4](https://doi.org/10.1016/0160-2896(81)90010-4)

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](https://doi.org/10.1016/0160-2896(81)90003-9)

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](https://doi.org/10.3102/1076998609332757)

Linn, R. L. (1976). *Fair test use in selection.* *Review of Educational Research, 43*, 139–161. [https://doi.org/10.3102/003465430403002139](https://doi.org/10.3102/003465430403002139)

Lippitt, R., & Waters, M. (1968). *Social behavior in children.* New York: Holt, Rinehart & Winston.

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](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](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](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](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](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](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](https://doi.org/10.1007/BF02289028)

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.

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](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](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](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](https://doi.org/10.1037/h0025105)

Lord, F. M. (1968a). *An empirical study of item-test regression*. *Psychometrika, 30*, 373–376. [https://doi.org/10.1007/BF02289501](https://doi.org/10.1007/BF02289501)

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

Lord, F. M. (1970). Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika, 35*, 43–50. [https://doi.org/10.1007/BF02290592](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](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](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](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](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](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](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](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.2011.11.002](https://doi.org/10.1016/j.cedpsych.2011.11.002)

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

Mazzaro, J., & von Davier, M. (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.carnegiknowledgenetwork.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.

McDonnell, 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/SICI1099-0984(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., & Fritzy, 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/s15327906mbrr0102_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/10.1177/0031644930590402002

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](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](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](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](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](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](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](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/10.1080/0020671.1975.10884742](https://doi.org/10.1080/0020671.1975.10884742)

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](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](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](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](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](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](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. (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](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](https://doi.org/10.1002/j.2333-8504.2008.tb02155.x)

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](https://doi.org/10.1007/978-1-4684-4172-7_2)

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. (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](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](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](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](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](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](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](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](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](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](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](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](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](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.

Swintont, 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](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](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](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/10.1007/10.1002/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). Bootstraping 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/10.1348/00711007X193957

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. Stepler (Eds.), *TIMSS 1999 technical report* (pp. 259–277). Chestnut Hill: Boston College.

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. Sottileare, A. C. Graesser, X. Hu, & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems* (pp. 169–178). Orlando: U.S. Army Research Laboratory.
Zechner, K., Bejar, I. I., & Hemat, R. (2007). Toward an understanding of the role of speech recognition in nonnative speech assessment (Research Report No. RR-07-02). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02044.x

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). 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

Zhang, M., & Deane, P. (2015). Process features in writing: Internal structure and incremental value over product features (Research Report No. RR-15-27). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12075

Zoccolotti, P., & Oltman, P. K. (1978). Field dependence and lateralization of verbal and configurational processing. Cortex, 14, 155–163. https://doi.org/10.1016/S0010-9452(78)80040-9

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 2.5 International License (http://creativecommons.org/licenses/by-nc/2.5/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.