CHAPTER 3

The Past as Preparation: Measurement, Public Policy and Implications for Access

SUSAN L. RIGNEY, DAVID E. WILEY, AND REBECCA KOPRIVA

The design, use and evaluation of today's achievement tests are the result of a century of development and debate. Substantial change has occurred in three broad areas: the tests and measurement models crafted to assess student learning, the ideas of how to verify the technical quality of these tests, particularly validity, and the public policy directing the use of achievement tests to evaluate students or schools. These three aspects of achievement testing are intimately intertwined. At various points, the technical aspects of testing appeared to be paramount, advancing procedures designed to evaluate defensibility but limiting the uses to which a test could be put. At other times public policy has driven innovation in test design and changes in the meaning and use of results. However, this legislation has often strained states and local governments who participate in the cost and resource allocation implications in order to comply. Each of these has impacted the notion and implementation of equity practices in academic assessments.

A Brief History of Large Scale Achievement Testing

Before World War I, educational testing was a fragmented enterprise, decentralized and the responsibility of individual teachers. College admissions were also fragmented. Individual colleges had separate tests until 1901, when the first College Board tests were administered to standardize college

admissions. These tests required applicants to complete essays in specific subject areas.

The first large scale test was the Army Alpha, an aptitude test created by Robert Yerkes and Carl Brigham to assess the qualifications of army recruits in WWI. Brigham is credited with the creation of the multiple choice item for this test (Carroll, 1990). In the early 1920s, Brigham also administered his own multiple-choice version of the Army Alpha test to Princeton freshmen. As a result, the College Board established a committee headed by Brigham to develop a test—to be called the Scholastic Aptitude Test (SAT)—which was used after 1926 by a wider group of schools for college admissions.

Multiple-choice tests—the so-called "new style" testing—were viewed as a great improvement over essay tests. Multiple-choice tests were valued for their fairness and objectivity because scoring was preset and not subject to the biases and preferences of the scorers. In 1933, IBM machines were used to score achievement tests for the New York State Regents and Providence, Rhode Island public schools. Mechanical scoring made scoring costs considerably lower than "subjectively" scored tests and large scale achievement testing arose in the 1930s and '40s to provide information over very broad grade-level content areas. By the 1950s multiple choice tests were firmly established as the preferred achievement test format because of their objectivity and efficiency. The introduction of automatic scoring machines permitted the implementation of testing programs on a scale that a review hailed as "previously been unthinkable" (Office of Technology Assessment, U.S. Congress, 1991). As a consequence an enormous jump in testing ensued and standardized multiple-choice achievement tests became widely used in U.S. schools from the 1950s through the 1980s. Successive waves of test-based reform were all intended to improve teaching and learning, e.g. the minimum competency tests of the 1970s and '80s. In general, results were reported as percentile ranks and individual student scores were interpreted relative to a national sample or norm group. Overall, remarkable growth of the testing industry during this period is reflected in the increased revenues from sales of commercial tests, which increased from about $40 million in 1960 to about $100 million in 1989.

In 1988, Cannell's criticism of norm-referenced achievement test results appeared. Often referred to as the Lake Wobegone Report, Cannell's obervation that all states were reporting results that exceeded the national average stoked controversy and many follow-up studies. The fall 1990 issue of *Educational Measurement Issues and Practice* was devoted entirely to the discussion. Shepard (1990) suggested "authentic assessments" as a possible alternative, noting that they may be especially warranted where score gains are not commensurate with other evaluations of student achievement. She

also discussed that authentic assessment tasks, if properly developed and implemented, held promise of measuring more complex knowledge and skills of students not usually possible with the multiple choice item type.

Widespread state exploration of more complex forms of assessment occurred in the early 1990s. The departure from multiple-choice-only format was motivated by several concerns in addition to those raised by Cannell. An increasingly urgent press for educational reform coupled with advances in research on learning and cognition sparked interest in test-driven curriculum and fostered use of item formats required students to demonstrate deeper levels of knowledge and skills. It also asked students to show their thinking and how they arrived at their answers, and their responses often took the form of written text or other types of open-ended responses. In general, there was dissatisfaction with multiple choice tests because they were viewed as emphasizing recall and basic application, rather than in- depth cognitive manipulation skills. Aschbacher (1991) argued that multiple-choice-only tests provided only limited information about the complex thinking and problem-solving skills of students and, as such, gave a seriously truncated view of the students' level of understanding. Further, Aschbacher suggested that these types of tests told teachers, parents and students that the more in-depth but untested skills are not as important as superficial knowledge and simple step application. Resnick (1987) noted that recent cognitive science findings challenged the notion of a progression from "basic" skills to "higher order" thinking skills and argued that all students "even non-English speakers, even the poor" were entitled to a rich curriculum designed to cultivate thinking skills as well as knowledge of basic facts. In order to foster this curriculum, assessments more suited to the nature of these abilities were required. As such, some state assessment units contemplated combining multiple choice items with a wide variety of innovative task formats, including open-ended items that could be answered in multiple ways, extended constructed response items or activities that included multiple steps and might require 15 to 45 minutes to complete. Some focused on revamping testing entirely, by creating portfolios that included a variety of student products.

In a study of state testing practices, Aschbacher (1991) found that, as of 1990, approximately half of the states were implementing, piloting or exploring the use of these new types of item formats, generally known as performance assessments[1]. While some states, such as California, New York, Vermont, and Kentucky, were actively working to integrate this approach into their state tests, the national enthusiasm for development of challenging content standards and new approaches to testing seemed to coalesce around Lauren Resnick and the New Standards Project. More than twenty states contributed funds and teacher expertise to the creation of

content standards based on research and practitioner knowledge about student learning in language arts, mathematics and science. Teachers from member states collaborated in the development of interesting performance tasks and field tested assessment components and instructional materials. Additionally, during the early to mid 1990s, consortiums through the Council of Chief State School Officers developed constructed response and performance activity item banks in science in social studies which were based on models similar to those being used by New Standards and some states. Finally, as association and state content standards were developed and used as benchmarks for the state tests, as the incorporation of innovative formats became more widespread, and as the focus shifted to grade span testing, the use of matrix-sampled assessments rose in popularity in an effort to handle questions of more in-depth curriculum coverage at the school rather than student level. This matrix model allowed states to utilize a broader range of time-consuming performance assessment items and evaluate a broader range of skills at the school level.

For individual and school accountability, the use of results from these item and testing types was hotly debated. The merits of assessments which in some way utilized these formats were sometimes presented as a trade-off between validity and reliability—more meaningful and authentic assessment tasks despite the challenges of hand-scoring and coverage that was sometimes less generalizable to the rest of the content domain. Generally the same expectations for validity and reliability were applied to performance-based assessments as for traditional test formats (Moss *et al*., 1992; Messick, 1995). Gong and Reidy (1996) discussed the tensions between the benefits of performance-based activities and the challenges of using the results for accountability purposes. Moss *et al*. (1992) asserted that student portfolios could generate important system level information in a comprehensive system of accountability to provide "an important supplement to the standardized sorts of performance assessment typically used at the system level and to suggest directions for curricular reform" (p. 20). Because performance tasks were sometimes memorable and therefore difficult to keep secure, new tasks often needed to be developed annually. This resulted in increased costs and made cross-year comparisons difficult when some of the more extended tasks or activities were used. Also, because complex tasks struggled to cover different topics within the domain, rarely did more than one task measure the same content. Thus, variability across tasks presented challenges in scaling and equating (Mehrens, 1992).

Besides the benefits to curriculum and the range of skills which could be measured, another important advantage of the open-ended item tasks was the increased involvement of content specialists in the development of assessments as well as the scoring and interpretation of results. Complex

items tasks were designed to elicit demonstrations of student understanding of essential curricular content and these tasks needed a scoring rubric that described qualitative differences across performance levels. Thus, these discussions yielded in-depth examination of what learning entails and what implies mastery and partial mastery. Further, it was the content experts who were best able to articulate the difference between, for instance, a "basic" and a "proficient" understanding of the concept. Over time, measurement and content experts learned how to build good rubrics apply consistent criteria to different tasks. This advance allowed the field to illustrate how students could demonstrate understanding of the same construct or skills in different ways. Good rubrics also communicated to teachers and students the central features of the concepts being tested and can be used as effective tools for classroom learning, and it was not uncommon for the rubrics to be taught and applied as part of daily lessons.

Because of the psychometric challenges that these new item types presented for large scale testing, the development activities generated lively discussions between curriculum specialists and psychometricians, and between measurement researchers, some of whom believed it was possible to improve validity while retaining a reasonable level of reliability across students. As opposed to earlier practices, for the first time, development of these types of items and pioneering test formats explicitly started from the concept to be tested and worked back to items and appropriate measurement models. In response to the psychometric challenges, over time scoring and scaling procedures were developed that supported valid and reliable results, particularly for certain types of tasks (Linn *et al*., 1995; Wiley, 1997; Young and Yoon, 1998; Young, 1998). The field also progressed in the types of criteria and scoring that were required to warrant comparable results across portfolios and other types of assessment activities. The good example of this advance was the underpinnings for the Advanced Placement arts exam from the Educational Testing Service, which allowed students to document achievement using a portfolio model that allowed different types of work products for different students.

Three issues persisted, however. Sometimes complex items tapped multiple targets, and measurement models based on an assumption of a unidimensional ability were not adequate at times. Second, practitioners discovered the difficulty of producing generalizable inferences among tests which used constructed tasks in some content areas (such as science) where certain task skill requirements were found to be less transferable. Finally, comparability from year to year among selected types of complex performance tasks was sometimes troublesome. Kentucky, for example, was forced to drop "performance events" (half-day hands-on activities) from its assessment program when results proved to be unstable from year to

year. Still, a few assessments that incorporated a range of open-ended items or tasks made adjustments and survived for a decade or more. These include the California Golden State science exam, the Kentucky writing portfolio, the Vermont mathematics portfolio and the New Standards Reference Exam' which was used as the statewide assessment in two states until 2005.

The use of complex constructed response items and formats declined or diminished in importance within the states' large scale academic assessment systems in response to the cost of continuous development and scoring, and the federal legislation requirements mandating testing at the student level for most grades (as compared with grade clusters). For the inclusion of the most complex types of activities and formats, the difficulty of developing some more complex performance item and test formats that were comparable from year to year, and related concerns regarding the technical qualities of specific assessments were highlighted. Recently, the pressure for quick release of assessment results for school accountability decisions has encouraged increased reliance on machine scoreable test formats.

Development of Measurement Theory Models

Measurement models—the theoretical underpinnings used to evaluate the stability, credibility, and meaningfulness of test structures and results—have changed dramatically in the last 100 years. Increased sophistication of measurement models has been aided by computers which permitted increasingly rapid and accurate completion of complex calculations. In this section the first part will summarize the major focus in educational testing to date, which has been to address the stability and generalizability of test inferences. The credibility of scores could be evaluated by some of the more sophisticated models identified in this part as well. The second part will outline salient advances associated with validating the "truthfulness", credibility and meaningfulness of the score inferences. This section will quickly review the history of how some major sets of models have developed. While most of the development of the theories in either part have not been tied to equity, *per se*, it would seem to be instructive to understand how the technical underpinnings are conceptualized. In this way advances which are meant to improve access can address the assumptions and other critical constraints associated with the technical framework of educational testing to date.

Reliability to Multidimensional Item Response Theory

Calculation of a reliability statistic (an estimate of the consistency of results over different test administrations or stability over students with like

abilities) was the central method of evaluating the quality of large scale achievement tests (for example, see Gulliksen, 1987, for more of a historical explanation). The dichotomous item scores of multiple choice tests required only simple indices of reliability. Spearman (1904, 1910) originated the theory of reliability by assessing the association between two alternative measures of the same thing. This led to the concept of test scores, or observed measurements, as true values plus error of measurement, which is the foundation of classical test theory (CTT). Although the original notion of measurement error in classical test theory terms came from discrepancies in scores based on parallel test forms, it is usually described today as random variation in scores caused by factors that are unrelated to the content tested. For instance, the theory of reliability developed, internal consistencies (or inconsistencies) among items within a test became an alternate way of conceiving of reliability (and error), as in Kuder and Richardson's statistic (1937).

Cronbach expanded the early notion of reliability by focusing on multiple sources of error and this resulted in his alpha coefficient (Cronbach, 1951) and ultimately in generalizability theory (Cronbach *et al*., 1972). Generalizability theory considers both test scores (performances) and the sources of variation (facets) that affect them, including differences among persons and test items. It attempts to explain variations in test performance by partitioning the variations among the facets. With this innovation, Cronbach and his colleagues initiated a change in psychometrics from emphasis on correlations to consideration of variance components and random effects ANOVA. This formulated and made practical multivariate generalizability theory, that, for the first time, allowed analysis and understanding of multiple abilities and their interrelationships and laid out effective ways of estimating universal (i.e. "true") scores for single and multiple abilities. Generalizability theory uses the same type of data modeled by CTT, but permits analysis of different facets of the test or test situation that may influence the consistency or stability of results. For example, one can examine the influence of different types of tasks, the number of raters, or different administration procedures. Results from these types of procedures could be used to improve the reliability of an existing test or to suggest desirable design elements for a new test.

The constraints imposed by the dichotomous item scores led researchers to question the use of common statistical models for test data which treated all variables as continuous and ignored the discrete nature of item scores. Item Response Theory (IRT) emerged as a solution. Tucker (1946) initially created the item characteristic curve (ICC), which expressed the probability of a correct item response as a function of ability, to adequately account for the dichotomous scores. This concept led to the contemporary way of

analyzing item functioning as a more definitive and invariant account of an item's difficulty and discrimination. Importantly, the invention of the item characteristic curve shifted the focus of measurement models away from the total test score to the characteristics of individual items.

Lord (1952) fully formalized the concept of the ICC into the "normal ogive" item response model that formally integrated test and item analysis, allowing test performance to be viewed as a summation of item performances in a common parametric framework. Coming from another direction, Rasch (1960) formulated a product-ratio model that led eventually to a logistic alternative to the normal ogive. These models were formally integrated by Birnbaum (1968) using new developments in mathematical statistics. Ultimately, logistic models were extended from dichotomous item responses to the graded score scales of constructed response items (e.g., Wilson, 1992). Within various assumptions, these models allow test and item analyses to be population invariant, yielding generalizations across populations. Further, the procedures make the equating of scores from multiple tests feasible even when the test forms are not formally parallel.

An important limitation to both the CTT and IRT models is that they provide only very limited diagnostic information. This is because they permit inference based only on the number of right/wrong responses. Pellegrino *et al*., (2001) point out that the recent development of some sophisticated measurement models make it possible to test or evaluate several aspects of student achievement simultaneously, or to consider variation in subsets of test items. These models could provide more detailed diagnostic information but they are seldom used. In general, the newer models have been developed to examine the relationship among responses within a carefully developed set of items designed to tease out the aspects of a concept that the student does not yet grasp.

Some of the models referred to by Pelligrino *et al*. included multi-dimensional IRT procedures. The notion that a single item within a test can measure diverse content has prompted this work where the logistic model for a single latent trait has been generalized to multiple latent traits. In 1980, the multi-component latent trait model was proposed (Whitely, 1980) to model the components underlying item response processes within an IRT framework. This model denotes the probability of success for a person on an individual item as the product of success probabilities for each of the underlying components. Due to the multiplicative, rather than additive, nature of this conjunctive, non-compensatory model, a deficit in proficiency on any of the components will lead to a smaller probability of getting the item correct. In this model, unless the person parameter is materially larger than the item parameter for all m components, the probability of a correct response to the item will be relatively low. There

is another class of multidimensional IRT models which represent the measurement components in a compensatory, or disjunctive, manner. Reckase (1985) proposed the model in which the undimensional trait level is replaced by a weighted sum of the composite traits. In this configuration a lack of one trait can be made up for by the presence of another. These advances provide insights that may help resolve some traditional problems in test theory and provide tools for scaling and equating tests in new ways. Further, while multidimensional IRT models do not directly address the issue of validity, these types of procedures begin to consider the degree to which the target and other influences are informing student performance. Differential influence on item scores could be an important indicator of some aspects of validity.

A Changing View of Test Validity

Within the testing community, the concept of validity has changed substantially in the past five decades. In 1950, Gulliksen's Theory of Mental Tests focused primarily on reliability and essentially viewed validity as the correlation between the target test and an external criterion, such as a parallel form. Cronbach and Meehl (1955) transformed the concept of measurement validation, reformulating it as analogous to theory validation as the philosophy of science understands it occurring in scientific fields. This perspective greatly expanded the array of evidence useful in establishing validity, but also attempted to bind it all together using the concept of a "nomological network" of potential empirical relationships as a way of encapsulating the theory of the measurement concept being validated. Textbooks from the mid-'70s defined validity as the degree to which a test or evaluation tool serves the purposes for which it is intended (e.g. Ahmann and Glock, 1975). Three types of validity were generally described: content validity, criterion-related validity, and construct validity, and practitioners were advised to collect data to support the type of validity consistent with the intended use of the test results.

Messick (1989) extended the Cronbach and Meehl framework by de-emphasizing the formal nature of the empirical network and by expanding the kinds of evidence that could be used in validation. Messick rejected reliance on a single kind of validity evidence tied to the test purpose and advanced the view of validity as a unitary concept that requires consideration of content, criterion, construct and consequences, with construct as the unifying force. He speculated that practitioners were reluctant to address the need for construct validity "perhaps because the process of construct validation seems complicated and vague" and because construct validation is viewed as a never-ending process. Messick continues

> "But just because a process is never-ending does not mean that it should not have a beginning. And a good beginning, at the least, is to attempt to discount plausible rival hypotheses about construct-irrelevant variance in the test. For example, a plausible rival hypothesis for a subject-matter achievement test is that it might, by virtue of its vocabulary level, be in part a reading comprehension test in disguise. . . . A variety of correlational, experimental, or logical approaches could be taken to render such rival hypotheses much less plausible" (p. 41).

It is important to recognize that, past and present, psychological views of learning and ability have strongly influenced the development of measurement models, the interpretation of assessment results and the validation of the inferences based on those results. It is often the case, however, that these viewpoints are present only as unstated assumptions in discussions of test development and analysis.

Pellegrino (1988) and others have argued that the traditional approaches to documenting construct validity, such as correlation and factor analysis, are based on trait theory rather than the discoveries of cognitive psychology that have documented the thinking strategies that individuals use in response to a wide variety of items. He suggests that the correlational patterns between test results that we have observed in the past may be less a function of a relationships between the tests themselves and more the result of consistent thinking strategies required across similar types of tests, so that validity might be better defined in terms of models of human thought, particularly its dynamics and limits. Moreover, he argues that instead of examining large numbers of items to determine the cognitive factors involved in a test, we should design the test to include items that systematically reflect the characteristics identified by cognitive science as relevant. Observation of the resulting pattern of performance on the test items could reveal more useful information about student understanding, or misunderstanding, of the construct.

Mislevy (2003) summarized the four major schools of thought that have influenced views of testing during the last century. First, trait psychology, which was associated with intelligence and aptitude testing at the start of the twentieth century, presumed that a mental trait or underlying ability is consistently evident across settings or activities and that observations across a variety of situations can provide evidence about a particular trait. Mental measurements focused one or more indicators correlated with competence in a domain. Assessments included multiple observations in the form of a diverse collection of items selected to "spread" students out along a performance distribution. This perspective is consistent with the

thin sampling of traditional achievement tests today. Mislevy notes that the psychological perspective shapes validation in terms of the questions that agree with or counter the inference we wish to make. Therefore, rival hypotheses for trait-based assessments seem to focus on the scope of the trait in question: do the items reflect the hypothesized relationships between performance and error or would another set of items suggest a different outcome? The second school of thought, the behaviorist perspective, regards knowledge as the "organized accumulation of stimulus–response associations, which serve as the components of skills." Measurement focuses on a target behavior from the perspective of the assessor. Success over multiple settings (tasks) is the goal. This is consistent with the notion in testing that small bits of knowledge, such as those represented by multiple-choice items, can be assembled to form a repertoire of complex behaviors. Assessments designed from this school of thought, such as the traditional criterion-referenced tests, estimate the probability of success in a domain to be commensurate with the amount of bits a student has mastered. Items are selected to represent essential knowledge in the domain and inferences at certain levels of percent correct suggest increasingly complex concepts of mastery.

Mislevy identifies the third school as the information processing perspective. This school emphasizes the procedures people use to acquire and strategically use knowledge in order to solve problems within the limits revealed by psychological experiments. The assessment design considers both task features and student performances. Inferences are made in terms of concepts and problem solving strategies rather than indirectly in terms of features of problems as an expert sees them. In this case, "the task a student solves is not the problem as the investigator poses it, but the problem as the student perceives it." (p. 18) Within a domain, tasks can still be grouped by features that are similar from the assessor's point of view, but the target of inference is the student's perception of features that make them similar from her point of view. Mislevy suggests that validation would consider an alternative explanation for results such as "Might some students be solving, say, purported spatial reasoning tasks using non-spatial strategies?" Fourth, is the sociocultural school which views knowledge as "conditioned and constrained" by the social context in which it is acquired and employed. Contextualizing assessment exercises decreases the assessor's control and increases burden of specifying the salient features of performances and performance situations. Validation questions to be raised and rejected may include "Do the relationships hold for some examinees but not others? Can the shared cultural experiences of low scoring students explain the results?"

Mislevy, Steinberg, and Almond's Evidence Centered Design (ECD) provides a blueprint for the operational elements of an assessment, that is includes the elements of identifying warrants, specifying procedures throughout test development and making sure that the evidence is generally commensurate with the inferences proposed by the test developers (e.g. 2003). The researchers maintain that

> "One cannot simply construct 'good tasks' in isolation . . . and hope that someone down the line will figure out 'how to score it.' One must design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them." (p. 2)

ECD describes multiple components and the ways in which they interact. One component is the task model. In addition to describing the materials presented to the student and the desired work products, it also includes features of the task as "a structure for understanding and controlling evidential variation." The researchers discuss the idea of controlled variation of task features that preserves the target construct provides comparability and supports a common inference despite planned differences in the appearance of the task. They say that the "family of tasks" from a specific template are not particularly meant to be interchangeable from a psychometric perspective (presumably where the item parameters remain constant), or necessarily from an access-based perspective. Rather, the model templates understand and control evidential changes and often intentionally manipulate the evidential value and item parameters.

As cognitive scientists have developed frameworks for the scientific understanding of learning, there has been increasing emphasis on the small-scale processes by which task performance capabilities are acquired and, consequently, on the fine-grained structures of the specific abilities of which they are constituted. As these processes and structures are elaborated, their combinations better represent the kinds of complex learning that takes place in schools. Individuals construct ever-evolving frameworks that juxtapose current states of ability with new experiences. The interaction of ability patterns with learning experiences results in new learning, e.g., transitions to new ability states. Current large scale test programs seem ill-equipped to address this view of learning, relying instead on measurement models grounded in a view of achievement as evidence of latent traits, and behaviorist relationships between knowledge of bits of information or skill and the inferences about mastery that they engender. As a consequence, test results represent "averages" of many abilities whose fine structure is

not taken into account. In addition, the interrelations which constitute "structure" are correlational, not contingent or prerequisite in nature. This may be the reason that it is so difficult to see most aggregate abilities as anything other than hazy collections with arbitrary boundaries.

Practically, indications of item content coverage in the test specifications, mostly internal review procedures during development, and the use of the same standardized administration procedures for all students were traditionally considered to be sufficient to ensure that the test was measuring what was intended and inferences could be made with confidence. In the '90s, Messick's work, the advances in learning, and the push for other types of items and formats brought a new focus to how validity might be enhanced and evaluated. This included reconsiderations of what validity means and new expectations of what constitutes adequate validity evidence. Investigations were undertaken which explored how to produce rigorous, independent judgments about the degree of alignment between content standards and tests, and the process became a popular type of validity evidence (for instance, see Hansche, 1998, for an explanation of alignment considerations). In 1999, the APA, AERA and NCME Testing Standards were published, which included a substantial redrafting of how validity is conceptualized and guidelines for how it might be demonstrated (AERA/APA/NCME, 1999). In this document, validity of score inferences were seen as a unitary idea under the broad umbrella of construct validity. The development of *a priori* validation arguments forms the basis for the evaluation of evidence completed to respond to the arguments. Mislevy (1995) and a few other measurement researchers (e.g., Wiley and Haertel, 1995; Shepard, 1990; Popham, 1994; Kopriva, 1999; Rigney and Pettit, 1995) called for more explicit links between what constructs and objectives the items are targeting, the processes that are used to develop, implement, and produce the scores on the assessments, and the evidence which was needed to demonstrate whether the items and tests were functioning as intended. While these researchers agreed on the foundation of good validity, their studies then differentially focused on different aspects of the measurement process in an effort to provide specific guidance about how score validity might be considered and verified in different situations, and for different purposes or students.

All of this work is forming a foundation for advancing how measurement of learning might be integrated into testing, and for raising the expectations about the kinds of validity evidence considered to be necessary and sufficient. Unfortunately, the progress in understanding how students learn is only slowly impacting how tests are constructed, and, other than independent alignment evaluations, most of the technical underpinnings for today's large-scale achievement tests continue to be indices of reliability.

The use of computers and the call for evidence of validity given the use of accommodations, however, are providing opportunities to rethink how items and tests are conceived, and to require improved verification that scores are reflecting intended constructs. Interestingly, besides the general technological advances in education, this latest push has been largely driven by recent federal policy.

Fifty Years of Federal Policy Regarding Equity and the Use of Achievement Testing

The Brown vs Board of Education decision in 1954 focused attention on equity in education and debate soon arose regarding racial inequality in testing and the adverse impact of some test-based decisions. The Civil Rights Act of 1964 advanced the federal commitment to equity in education and in 1974 Lau vs Nichols spelled out the educational rights of language minority students. In 1965, Title I of the Elementary and Secondary Education Act (ESEA) initiated federal funding to support educational improvement for economically disadvantaged students. Although education remained the responsibility of each state, federal funds through ESEA were initiated and continue to be distributed as a national commitment to the education of children with disadvantages, including high poverty students, English language learners, students with disabilities and migrant students. Jennings (2000) summarizes the historical shifts in political support for the ESEA and associated changes in funding through the early 1990s when political leaders from both parties called for increased education funding accompanied by a shift in the accountability focus from consideration of "inputs" such as funding to improved student "outcomes" as represented by achievement test results.

ESEA: Full Inclusion, Disaggregated Reporting, and Beyond

The movement to standards-based instruction and assessment in the early 1990s represented a significant paradigm shift for educators and testing professionals. The 1994 reauthorization of the ESEA, known as the Improving America's Schools Act (IASA), produced profound shifts in both testing policy and practice. First, under Title I of IASA, all states were required to develop a system of challenging academic content standards and assessments with clearly defined performance standards. Moreover, these standards applied to all students. The shift from reliance on norm-referenced tests to criterion-referenced tests was made to change the logic of test development as well as how test results would be interpreted. Each state was now faced with defining not only the essential content to be tested but also the target performance on the test that represented "how good is good enough." The challenge of test design became the creation of items

that were supposed to represent the content critical for all students rather than simply a set of items that could sort students along a normal curve. Second, state assessments were required to include all students, including students with disabilities and students with limited English proficiency. For ELL students[2], states were advised that:

> "The appropriate form of assessment might be assessing the student orally or in writing in his or her native language; providing accommodations such as a bilingual dictionary, extra time, or simplified directions; using an assessment that has been stripped of non-essential language complexity; or administering an English language assessment orally." (U.S. Department of Education, 1999b)

Third, the assessment results were to be disaggregated and reported publicly for the following groups: all students, gender, major racial/ethnic groups, ELL, students with disabilities, economically disadvantaged, and migrant. The publication of disaggregated assessment results was a significant change in testing practice, certainly for demonstrating achievement at the federal level and for most other purposes as well. The requirement was intended to ensure that all students in schools receiving Title I funds received the instruction they needed to reach the achievement standard, including additional support as needed. Public reporting by disaggregated subgroups was the method of holding schools accountable for services provided under Title I.

Despite the changes associated with the adoption of standards-based assessments, the most profound change was undoubtedly the federal requirement of full inclusion. The commercial norm-referenced achievement tests routinely used as measures of student and school accomplishments for more than 50 years, were norm-referenced and did not emphasize the inclusion of most English language learners in the norm group. Since most states relied on these tests as the basis of their assessment program prior to 2001, ELLs were generally exempted from state and district testing. In 1998 all but nine states permitted most ELLs to be exempted from statewide assessments, usually on the basis of the amount of time in the U.S. (tewnty-seven assessments) or length of time in a school program where they learned English (Heubert and Hauser, 1999).

The inclusion of all students, including those that had previously been exempted from testing by state directive, was required in the 2000–01 school year. As the deadline approached and it became evident that few states were implementing this requirement, the U.S. Department of Education released a policy letter that included the following direction:

> State policies must guarantee that each LEP student is included in the State assessment system. LEP students are to be provided an individualized determination of the most appropriate language and form of assessment for that student, based on English language proficiency, native language proficiency, language and format of their current instructional program, or other relevant factors. Whether an LEP student should be tested with the State assessment, the State assessment with accommodations, or (to the extent practicable) a native language assessment will depend on which assessment most validly and reliably measures her or his knowledge and skills. In no instance may a State assess an LEP student against content or performance standards less rigorous or less demanding than the standards applicable to all other students. Accordingly, a blanket State exemption policy for LEP students for Title I purposes, whether permissive or mandatory based on time in U.S. schools or time in English instruction, would not meet the Title I requirements (Cohen, 2000).

Despite what appeared to be slow progress in the implementation of fully inclusive state assessments from 1994 to 2001, it was within the last two years of this period that virtually all states changed the statutes and procedures that had previously required or permitted the exemption of most ELL students. However, by January 2001, only eleven states had received full approval for their assessment system as required under IASA. Many of the remaining states were criticized by the U.S. Department of Education for failure to include students with disabilities and limited English proficiency in the assessment system. All in all, the "guidelines and criteria for the inclusion of LEP students in assessment were inconsistent across and within states (US Department of Education, 1999b)."

The U.S. Department of Education's peer review of state assessment systems under the 2001 reauthorization of ESEA, No Child Left Behind (NCLB), moved beyond simple participation and reporting requirements to emphasize the use of accommodations or alternate test formats for students with disabilities and for English language learners, and to require that adequate documentation of validity as well as reliability accompany their use. For instance, the peer review included questions about the validity of scores based on accommodated test administration, including changes to the test format such as translations or use of simplified English. In addition the review considered whether states actually monitor the delivery of accommodations on test day. For alternate test formats, the federal guidelines specified that the technical underpinnings of these tests or forms be similar in rigor to what was expected for the mainstream tests. In the

initial review, few states provided such evidence for either special needs students taking the mainstream test or for those who were tested under in an alternate format. This raised concerns about whether students were actually receiving the accommodations or other options needed to "level the playing field", thereby permitting them to demonstrate their reading and math skills independently of a disability or limited English proficiency. Most recently, NCLB has required publicly released reports of school accountability based on subgroup test results as a critical strategy for ensuring that all students, including English language learners, have equal access to the states' achievement standards.

Other Federal Influences

Other than the ESEA authorizations, the other federal law that has had a direct influence on states' large scale achievement testing is the Individuals with Disabilities Education Act of 1997 (IDEA). IDEA required all states and districts to include all students with disabilities in their testing programs. For most students with disabilities, this was to be accomplished by using accommodations specified by their local Individual Education Plan (IEP) teams that were consistent with what was being used in the classroom. In addition, implementation of alternate assessments for all statewide and district-wide tests were required by July 1, 2001 for those students that were not able to participate in state- and district-wide tests even with accommodations. Although most states complied by implementing alternate assessments for their most cognitively disabled students, the result from these assessments generally represented a less rigorous achievement standard than the general assessment. Further, the technical underpinnings were sometimes questionable. For students receiving accommodations, one results from the IDEA legislation was that the number of possible accommodations mushroomed, and states and/or districts were left to sort out which accommodations would threaten the validity of the tests and which would not. Unfortunately, several reports have shown that, over states and testing programs, there was surprisingly little agreement on which accommodations were or were not allowed, and which accommodations were being assigned to which students. So, while these students with disabilities have recently been included in testing with some accommodation help, questions remained about the adequacy of the accommodations, the consistency of implementation across students, and the effects the accommodation use was having on the test results.

During this time, and particularly once NCLB was authorized, accommodating English language learners often involved adapting the guidelines and accommodation lists developed for students with disabilities. Initially, many of the accommodation options were not appropriate, and since there

was no IEP team to assign accommodations to students, with little or no training, the decision-making was usually left to the student's teacher or to other staff identified by the state or district. Over the last few years there have been efforts made to determine which accommodations are suitable for ELL students, to improve how the students are assigned accommodations, and how auditing and oversight might be handled (see later chapters for more of an explanation). As noted above, ESEA has taken the lead in monitoring and evaluating the use of these options for both English learners and for students with disabilities.

Federal legislation mandating the creation of The National Assessment of Educational Progress (NAEP) has had an impact on equity in academic assessments as well. NAEP was introduced in 1969 to provide a national view of student achievement over time based on a nationally representative sample of students. NAEP, which samples students across the nation in a few grades per content area every few years, has not directly influenced state assessments. Rather, it has primarily served as a model, providing an example of rigorous item types, sophisticated matrix design, and complex statistical modeling and analysis of results. NAEP permitted the exclusion of most ELL students until 1996 and most states followed NAEP's example. Ironically, an evaluation of NAEP reported that while many students with disabilities and English language learner students had been excluded from previous NAEP assessments, many of them were actually capable of participating in the assessment (Mazzeo *et al*., 1996). This was found to be true particularly if certain types of adaptations and accommodations were made available to the students. To prepare for a possible change in policy, the NAEP program pilot tested accommodations in 1996 for special-needs students, including English language learners, and revised inclusion criteria to make them clearer and promote uniform application across participating states. By 2002, NAEP provided accommodations on all assessments for both English learners and students with disabilities, although some constraints for inclusion still remained.

Finally, during this period, commercial publishers responded to NAEP's decisions and state pressures and increased the number of students with disabilities and limited English proficiency in their norming and validation samples to represent the full student population (see CTB McGraw-Hill, 2005 for a test publisher's discussion of the impact of accommodations on the interpretation of test results). However, few of these publishers disaggregated their technical analyses of their tests to ensure that the assessments were appropriate for these subsets of students as well as the majority population. This left the test consumers and measurement field in a position where some of the technical work seems to be lacking.

The two requirements by NCLB—requiring adequate evidence of score validity for students who use accommodations, and emphasizing the technical expectations of alternate formats—is encouraging the state testing programs to expand how they typically complete the technical work necessary to document that the tests results are stable, credible and meaningful. There is a tension. On one hand, it makes sense that states, who are responsible for the teaching of all students, should be responsible for defending the technical adequacy of their assessments which are meant to measure the students' learning. This includes demonstrating that all students are being held to the same standards, regardless of how they are tested. On the other hand, new approaches and investigations are needed to adequately demonstrate that there is adequate technical quality for the implicated students because the foundation of technical procedures is not well known for students who are tested under varying conditions. These technical advances are taking time to create or adapt from other literature, and the cost implications of procedural development and additional analyses is sometimes sizable. The question becomes, why weren't these methodological advances integrated into testing systems until now? While it is painful and time consuming for states, the federal government has found itself, in this situation, on the cutting edge of the measurement field.

Some researchers, including Solano-Flores and Mi (2006), Abedi (2006) and Kopriva and others (Emick and Kopriva, 2006, 2007; Kopriva *et al*., 2007) argue that the current efforts do not yet reliably translate to equitable and valid outcomes for some English language learners. These researchers maintain that scores for students who can fully access the test items represent information about the targeted constructs, but for students who can not fully access the items, the scores still represent a mix of information about the intended information and information about irrelevant factors unrelated to the targets. Some would argue that when test results are used for accountability purposes, this is tolerable because at the school level it results in a cautious application of accountability consequences that favor students. Under NCLB, for example, depressed results for ELLs could result in the opportunity to attend a "better" school or gain tutoring services. However, since it is clear that the bias begins at the item level for these students, as items are aggregated to produce a test score, the results represent a distortion that potentially masks where individual content problems lie (and where they don't). Further, because of the misunderstanding of student needs, there is a very real possibility that depressed scores at the ELL subgroup level leads to an inequitable allocation of school resources and valuable instructional time. It appears that improving how the achievement of ELLs is measured in states and schools may be a useful utilization of funds to improve the schooling for this population.

Building Assessments to be More Accessible for ELLs

Wiley (1990) distinguishes between the target construct or *intent* of a measurement and the *ancillary* abilities of individual students necessary for a successful task performance. For example, test taking abilities are well known to contribute to performance on multiple-choice tests such as the SAT. Individual differences in such skills constitute part of the "reliable" variation in scores on any test. However, such abilities are not part of the measurement intent of these tests; they detract from validity or, to put it the other way around, they add to the invalidity of a test. In statistical terms, both the intent of measurement and the ancillary abilities are dimensions of measurement. He asserts that ancillary abilities, their distribution over test tasks and their relations—both among themselves and with the characteristics intended to be measured—are critical to an adequate conception of test validity. From this perspective, test validity must be assessed by examining all of the dimensions which contribute to test performance.

Certainly, accessibility is a factor of both reliability and validity. However, it seems that improving the validity in score inferences for students with identified challenges is of particular interest because, without confidence that the scores reflect the intended meaning, consistency in measurement is rather meaningless.

The heart of intended measurement is a clear understanding of what is being measured. This includes explicit statements about targeted knowledge and skills not only at the test level but at the item level as item scores are generally aggregated to produce test scores. As noted above, unlike random error which assumes a cumulative net effect of about 0, systematic error over items increasingly adds distortion to test scores. Therefore, when accessibility is a question, it seems prudent to minimize the effect of irrelevant systematic influences by focusing on the components of testing that impact the item level, including item writing, internal test development procedures, implementation, etc. The following section will briefly outline some of the aspects of accessible test-making.

Conceptual Underpinnings

In general, today's item and test performance models focus on how test takers interact with items, *under the assumption* that the interaction is only or mostly about the targeted content of the item and the student's knowledge in response to the item's problem.

Beyond this, the test taker/item interaction seems to be a neglected area of focus. When items or tests are inaccessible for some students, however, it can be said that the effects of irrelevant and ancillary factors *interfere* with the measurement of targeted knowledge and skills. In all cases, some

amount of irrelevant information always accompanies the communication of any item to a test taker. In most cases, the function of the irrelevant information, such as the fact that the item's requirements are written on a page in English, is to *facilitate* the communication of the item's problem, not inhibit it.

It is suggested that the interference, when it occurs, is actually a function of aspects of the person/item interaction, when irrelevant characteristics in items interact with sub-par skills of particular students.

Recent advances in cognition and measurement provide a basis for thinking how students approach, address, integrate and retrieve concepts and skills. Further, research supports that students move through identified sequences in different ways and at different rates, depending on a multitude of attendant individual and environmental factors. However, much of the work on task processing and learning seems to have been focused on qualities of the tasks and student competence regarding the desired task targets, rather than on identifying differential ancillary characteristics of individual students and how these interact with processing in a variety of tasks. For instance, Embretson and Reise (2000) identifies task-specific cognitive sub-processes in tasks and models how students respond to the different sub-processes so that their varying scores are a function of their degree of mastery in the target abilities. Lohman and Bosma (2002) point out that both experimental/cognitive and differential/measurement psychologists frequently array their data in a person by task matrix, and that both sets of psychologists have tended to emphasize the main effects in this matrix. While experimental/cognitive psychologists emphasize differences among tasks/treatments and differential/measurement psychologists emphasize differences among persons, both desire to minimize the interaction between persons and tasks without focusing on the quality of the differential interactions *per se*.

There is some work that attempts to explore the person/task interactive space and consider what is causing the differences across students. In a major review, Pelligrino *et al*., (1999) focused on intelligence and aptitude tasks, explaining how the cognitive components approach worked to develop differential models of task performance by exploring components of performance that varied across individual task takers. The approach assessed performance strategies, executive routines, and how targeted declarative and procedural knowledge interacted with the varying processing capabilities of task takers.

Some researchers explore the targeted and construct irrelevant aspects of the task/construct space while generalizing across students. Snow and Lohman (1993) appear to draw a distinction between component skills and strategy adoption, suggesting the separability of perception, memory,

verbal and special abilities (as well as strategy) from the targeted construct. Glaser and Baxter (2002), among others, define a content-process domain space for school science within which targeted constructs can be classified and defined in terms of task types. The four quadrants of the space (content—rich to lean, process—constrained to open) provide a context for differentiating targeted construct aspects of task performance from construct irrelevant aspects, if such item characteristic differentiations are specified. Mislevy *et al*.'s (2003) ECD provides a basic architecture for developing assessments that permit variations while preserving the inferences we wish to draw regarding student competence in the domain. However, in this document and elsewhere, while the researchers have spelled out the key elements which are addressed in assessment development, how target task elements may change, and what these two elements mean for the test development architecture, the work does not focus on how student variations due to irrelevant influences may impact the testing work. This seems to be the key area of focus for accessibility purposes.

It seems that work such as Snow and Lohman, Glaser and Baxter, and Mislevy *et al*., open the door for developing more individually tailored items and assessment approaches keyed to measuring the same target abilities while minimizing the effects of the irrelevant factors through active manipulation of their environments. In this case, the focus would be on understanding particularly how items and tests need to be designed and developed to maximize the measurement of targeted knowledge in the person/task interaction space for students with certain challenges. In most cases, the objective is to develop assessments that can defensibly yield common inferences and comparable scores across persons with and without sub-par ancillary challenges. The work also involves understanding the types of evidence that will be needed to make sustainable arguments about comparability and common inferences.

For instance, studying accessibility (and inaccessibility) investigates under what conditions target skills are properly conveyed in the person/item interaction, and when communication about targeted information becomes systematically contaminated, misunderstood, or distorted. This contamination or distortion is known as systematic error. A detailed focus within the encounter would revolve around how the same assessment tasks differ from individual to individual. A less detailed approach could look more broadly at groups of students sharing certain sub-par skill levels of key variables, and groups of items that share certain active characteristics.

The argument for developing accessible testing systems is based on the essential notion that variations in procedures and materials could be designed to produce common inferences. The grounds for common test inferences are traditionally found in a *procedural argument*: common

content in items and a common approach for synthesizing and summarizing items and response data over items. The latter part of this procedural argument required standardized conditions of observation as a key aspect of synthesizing item data. However, based on developments in measurement and cognitive science, one can make an alternative conceptual argument for common inferences. As in the procedural argument, the measurement of common substantive content is important. But rather than requiring standardized conditions of observation, the *conceptual argument* can be built on evidencing appropriate interrelationships between target inferences, the knowledge and skills of interest, the properties of tasks or items designed to elicit the observations, student characteristics that impact testing and items, necessary observations, and the assessment situations where students interact with assessment requests. This approach suggests that data may be collected under alternate conditions (Mislevy, 1995). By minimizing the influence of irrelevant input on student performance without adding significant additional sources of error test developers can increase the validity of the task inference without sacrificing reliability. At its crux, Kopriva (1999b) suggests that, when common inferences from a robust conceptual argument are applied to assessment results for which there is sufficient evidence of minimizing systematic error across testing procedures and materials, this should provide the grounds and be the basis for determining the validity and comparability of scores.

Whether access is considered from a detailed or more superficial level, two directions for further work seem to be central. First, there needs to be work completed which is aimed at ameliorating the interference of inaccessible stimuli in order to improve the quality of the students' performances so they more accurately reflect their targeted knowledge and skills. Second, it seems prudent to develop some "barometers" or other algorithm-based estimates which would calculate the level of access that is available for different students in different testing systems. The more detailed estimates can also be used to adapt future flexibly built computer-based testing systems in order to maximize the ability of individual students to receive particular items and formats that are aligned to their needs and strengths.

Developing Accessible Tests

The following chapters in the book will explain in detail how tests might address the issue of accessibility throughout the processes of item/test development and implementation. As the reader attends to the information in the various chapters it will be important to keep in mind that the person/item access interactions seem to occur primarily at three intervention points:

1. For the interaction between the student's targeted knowledge and skills and the item's request for targeted information to successfully begin, the student must have sufficient access to how the meaning and the requirements are conveyed in the task.
2. For the student to initiate and sustain problem solving activities relative to the task requirements, students must be able to access their procedural skills and other ancillary content, and have the tools necessary to implement the activities. (The ancillary content and skills would be considered construct irrelevant and within the domain of access to the extent that they are not part of the target requirements.)
3. For the student to represent their solution to the task requirements, students must be able to access their representation skills commensurate with the representation constraints in the particular task.

Thus, the first intervention point for access could be seen as understanding what the task is asking. Lack of prior instruction, or a confusing graphic that is part of the task, or reading skills are examples of how access might be blocked for some students at this stage. The second point focuses on the students' access to tools or procedures or language which would help them solve the item problem but which are not a part of the targeted objective. For instance, if the solution requires the access to language about wintertime in a cold climate to be able to compose a poem, a student's inexperience with snow, for instance, may be the cause of failure to solve the problem. Third is the generation of a response to the item. A student who understands the science but lacks the academic language skill or learning opportunities to differentiate properly between possible answers in a multiple choice item may not be able to convey his solution properly. (It can also be suggested that problematic options can hinder the students' problem solving, if the argument can be made that eliminating possible answers is a viable approach to thinking through the problem.) All of these access barriers may be classified as characteristics of the task, and should be considered when identifying challenges and strengths for the student. In many cases elements at each of the intervention points can be varied to remove or reduce barriers that prevent the individual who understands the concept from generating a correct response.

It is argued here that two models or templates are at the heart of any test which purports to be access-based. The first is the *Access Specification Package* explained in the next chapter. This set of specifications is broad-based and a macro-indicator, specifying how access should be considered throughout test development. It calls for evidence of access at the item level (including item writing, reviewing, and implementation procedures); at the

testing system level (for instance, what accommodations are allowed and are these a sufficient, defensible, and parsimonious set?); as the testing system expectations interact with school implementation systems (e.g. are the proper accommodations assigned to the proper students?); and at the consumer level (are the scoring and analytic procedures adequate?). It is designed to be used to guide development and also to evaluate the level of access that can be assumed if components are implemented as planned.

Second, the nucleus of the *Access Specifications Package* for ELLs is the item writing and materials development. The model which necessarily guides this development is the *Access-Based Item Template*. The template is explained in Chapter 5 and is relevant particularly for activities proposed in Chapters 6–8. For comparability purposes, the target and other information required in the template core is of a finer grain-size than those specified by Mislevy and others. This is because it includes an analysis of ancillary characteristics of each item as well as specifying the substantive target. It is argued that this detailed level of specification is required for score equivalence to be defended. The goal was to construct a template with reasonable demands on the item writer, but one that is particular enough to specify key characteristics of students who would appropriately benefit without providing an unequal advantage for some students.

Estimating Levels of Access

The types of multidimensional models of item performance identified by Whitely, Reckase, and others would seem to be useful as a starting point for modeling how access factors interact with one another in an item/testing process for a particular student (Kopriva *et al*., 2004). In some ways, access is a compensatory exercise. For instance, a student with a low level of reading skill may be able to compensate for that deficit if the item includes a visual that makes the meaning of the question clear and the student can decode the visual in a meaningful way. However, when items and tests are not accessible enough, the overall effect seems to be conjunctive as well. In this case, inaccessibility appears to have an inversely related multiplicative effect on target performance, especially when certain non-target abilities are seriously below the threshold level required for the student to properly interact with the tasks. All in all, Wiley *et al*., suggest that, for some students, it is necessary to model score performance as a function of access as well as target ability, and that access factors in this approach need to be modeled to address both conjunctive and compensatory elements.

These researchers and others have completed some preliminary work related to how item and test performance might be modeled. They suggest that, to model access for a testing program with certain kinds of forms and accommodation characteristics, overall levels of accessibility could be

modeled for subgroups of students, for instance students at different ELP levels. To date, key student and item/testing characteristics have been identified and judgments have been made about the kinds of relationships between item/test and person variables for various subgroups of students. A preliminary prototype modeled some of the item access-variables, but person variables were not at a sufficient level of detail and test conditions were not considered (Wei *et al*., 2005). A related study (Hipolito-Delgado and Kopriva, 2006) suggested that ameliorating some of these key testing variables seems to differentiate performance for students with particular needs, and it appears that this work, as well as other research, can be used to advance the model development. Some introductory work has begun on formulating a differential item functioning statistic as well that may be able to minimize a key problem for ELLs in how DIF methods are currently conceived (see Chapter 12 for more of an explanation about DIF).

For research purposes and for use with future access-based electronic item/test procedural banks, the authors and other researchers have started to formulate what a person/item model might look like. The model would include not only relevant item and person data but also information about what testing conditions would be appropriate for a student with particular needs and strengths. They observed that student and item/condition matches seem to follow a series of procedures and rules that include

1. the identification of the target and ancillary factors evoked by particular items/conditions across students to form a target and access profile vector for each item/condition combination;
2 the identification of interactive rules within the item profile vector for specified factors across students with different student profiles;
3 the identification of factors in student profiles across items/conditions to form a vector profile by student;
4 the weighting and interaction of target and ancillary factors within student vectors to provide prioritized matching expectations that differ by student.

Identification of target and non-target factors and interactive rules for persons and items/procedures includes determining the presence or absence of specific components in test components or in students. The model also needs to specify the threshold amounts of each component and weights across components. In part, interactive item rules would both explain when the item requirements or problem solving mechanisms can be accessed by more than one factor or when the presence of one factor could lower the threshold for another factor. For students, some factors may be necessary

(although not sufficient), or certain alternative student profiles may apply to specific item configurations.

In general the research approach is to create a target/access profile vector for each student that will be used across items/condition combinations, and an item/conditions target/access vector for each item/condition that explains what factors are evoked in the particular testing situation. In each case, as noted above, the student and item vectors are "customized" by additional rules of compensatory and conjunctive interaction and weighting that differ over students and items, respectively. Once these profiles are identified and the rules programmed by student or item, it appears that the matching can be completed electronically. A recently completed computer-based method to assign appropriate accommodations to particular students has successfully tested out some of these objectives (for instance, see Kopriva *et al*., 2006b, and Chapter 10 for more information). It is likely that this system could be used to move towards such a person-specific access model. One application may be that, when items are available electronically, algorithms can be built to identify the best match among like items and testing procedures can be found for each student for each identified targeted construct element that a test is intending to measure.

This chapter represents a great breadth of material about forces which helped influence what today's tests look like. Over the last 100 years, it surveys how specifications of valued academic content, theories of learning and cognitive acquisition, consumer appetite, and public policy, have specified, changed, and constrained what achievement tests focused on and why. Over time, technical models of measurement were developed and adapted from other literature to specify how assessments might be structured, and what components were needed or assumed in order to defend scores based on concepts of stability, consistency, generalizability, "true" mastery, and meaningfulness. However, not until recently has there been a more widespread acknowledgement that ancillary components as well as targeted objectives impact the scores of test takers. As this chapter has outlined, it has become clearer that, because of the nature of how today's standardized tests request and communicate information, the language, cultural, and contextual challenges of English language learners and others seem to form a web of small blockages to access that pervade tests and test items. Past a point, these interferences seem to be impervious to traditional equity approaches, such as readability and bias reviews, and to easy post-development accommodation attempts.

It will be a substantive challenge to integrate current research on learning, emerging concepts about test design, advances in multidimensional models, and work that specifies components of tests and students, with a growing understanding of how test takers interrelate with the business of

academic testing. This is particularly the case for students like English language learners who have been regularly excluded from large scale academic testing until just recently, and whose challenges (and strengths) are complex and multi-faceted. However, given the technological advances, and groundwork in understanding some key elements of student/testing interactions for this population, the rest of the book is designed to explain how this challenge might be advanced.

Endnotes

1. "Performance assessment" typically referred to the use of an open testing format (such as portfolios) or incorporation of some type of open-ended items or prompts on a standardized achievement test that requires a student product generated in response to a test question or task. Products from the open testing format or constructed responses on a standardized exam are subsequently evaluated by means of a scoring criterion consistently applied.
2. English language learner (ELL) is the preferred term to indicate a student who is not yet fluent in English; however, federal statues and publications employ the term limited English proficient (LEP) to refer to these students. These terms are being used interchangeably in the book. For ease of reading, ELL will be used here although the statutes use LEP.