Review of Research in Education

http://rre.aera.net

What Counts as Evidence of Educational Achievement? The Role of Constructs in the Pursuit of Equity in Assessment

Dylan Wiliam

REVIEW OF RESEARCH IN EDUCATION 2010 34: 254
DOI: 10.3102/0091732X09351544

The online version of this article can be found at:
http://rre.sagepub.com/content/34/1/254

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Chapter 8

What Counts as Evidence of Educational Achievement? The Role of Constructs in the Pursuit of Equity in Assessment

Dylan Wiliam
Institute of Education, University of London

If what students learned as a result of the instructional practices of teachers were predictable, then all forms of assessment would be unnecessary; student achievement could be determined simply by inventorying their educational experiences. However, because what is learned by students is not related in any simple way to what they have been taught, assessment is a central—perhaps even *the* central—process in education. At the very least, assessment is integral to effective instruction.

At first sight, it appears that assessment should be relatively uncontested. Everyone—parents, teachers, employers, the wider community that supports public education through taxes, and the students themselves—just wants to know what it is that students have learned. However, two difficulties emerge immediately. The first is that by its very nature assessment *reduces ambiguity*. The fifth-grade mathematics standard for many states requires students to be able to compare two fractions to find the larger, but when we assess, we have to decide which pairs of fractions should be included and which should not. This may be done explicitly, through a formal process of construct definition (see below), which lays out clearly what should be included and what should be excluded from the assessment, or more commonly, it may be done through some less formal process, involving a judgment of what is appropriate, given that this standard is intended for fifth-grade students—as William Angoff (1974) remarked, "lurking behind the criterion-referenced evaluation, perhaps even responsible for it, is the norm-referenced evaluation" (p. 4).

In fact, the choice of the fractions to be compared makes a huge difference to the rate of student success, even if we restrict the domain to fractions where both numerator and denominator are less than 10. In fifth grade, where the fractions have equal

Review of Research in Education
March 2010, Vol. 34, pp. 254-284
DOI: 10.3102/0091732X09351544

denominators, more than 90% of students are likely to be able to succeed, although fractions with unequal denominators are likely to present more of a challenge (Hart, 1981; Vinner, 1997). However, where the fractions have unequal denominators but equal numerators (e.g., $\frac{5}{7}$ and $\frac{5}{9}$), then fewer than one student in five is able to answer correctly (Hart, 1981). If a requirement to assess something as apparently precise as "can compare two fractions to identify the larger" can produce such different interpretations, even where the domain is restricted to single-digit numbers, then it is hardly surprising that even when different stakeholders agree about the importance of the material to be assessed, the assessments themselves become vigorously contested.

The second difficulty arises because assessments are *representational* rather than *literal* technologies (Hanson, 1993). When we assess students, we are never interested in how well they do on the actual items on which they were assessed; we are interested in how we can generalize beyond the behaviors observed on the assessment (Nuttall, 1987). The desired generalizations may be in terms of future performance in higher education, as is the case with the College Board's SAT; items similar to the ones on which the students were assessed, as in the case of the fractions example above; or even, in the case of a spelling test, whether the students will recall correctly tomorrow what they could do today. Calfee, Lau, and Sutter (1983) suggest that the ability of a soldier to strip down and reassemble a rifle may be a valuable skill in its own right, but as they acknowledge, even here, the reason for practicing the drill is that the skills become so automated that they can be executed in less ideal settings (e.g., at night, knee-deep in a swamp, under fire).

As Nuttall (1987) suggests, "The fidelity of the inference drawn from the responses to the assessment is what is usually called the *validity* of the assessment" (p. 110), and in this chapter, I will trace one strand in the development of the theory of validity—the increasing importance attached to the role of constructs in validating educational assessments—and show how, ensuring that the primary focus is on the construct of interest, rather than the assessment itself, can bring some greater clarity to a number of debates, especially in the area of equity in assessment. To illustrate how a focus on the construct of interest can clarify the debate, I discuss, in some detail, three particular arenas of assessment:

- Testing for admission to higher education
- The rise, and fall, of measures involving constructed-response items
- The assessment of students with special educational needs

Within each of these areas, I explore the interplay of issues of technical adequacy with equity and show how attention that has in the past been directed at the adequacy of the assessments might be more fruitfully directed toward the construct of interest. In other words, I argue that attention should be shifted from how well we measure something to what it is that we are measuring. By clearly separating the values issues—what we should be assessing—from the technical issues—how well we are

assessing—we allow greater public engagement in the debate about what should be assessed (because those lacking the necessary technical expertise are not excluded from the debate). Although this is not, in itself, a guarantee of equity, it does, I believe, create the possibility for a wider debate in which previously underrepresented voices can be heard. Greater clarification about the construct of interest may also increase the technical quality of the assessments themselves.

THE ROLE OF CONSTRUCTS IN VALIDITY THEORY

The idea that validity should be considered a property of inferences, rather than of assessments, has developed slowly over the past century. In early writings about the validity of educational assessments, validity was defined as a property of an assessment. The most common definition was that an assessment was valid to the extent that it assessed what it purported to assess (Garrett, 1937) and this definition is still in widespread use (Brasel, Bragg, Simpson, & Weigelt, 2004; Cohen, Mannion, & Morrison, 2004). The problem with such a definition is that an assessment does not actually purport to do anything—the purporting is done by those who claim that a specific assessment outcome has a specific meaning—a test tests just what a test tests. Of course, tests come with labels, which convey implicit or explicit claims about what the test does, in fact, test, but, as Kelley (1927) pointed out, this can be highly misleading. Two tests with very different labels may be assessing very similar things, and two tests with the same label may be assessing very different things. Hence, as Nuttall (1987) points out, "In practice, an assessment does not have a single validity; it can have many according to its different uses and the different kinds of inference made, in other words, according to the universe of generalization" (p. 110). A test can be valid for some purposes, but not others. Performance in mathematics tests at age 16 can be shown to be good predictors of mathematical performance at age 18 (Wiliam, Brown, Kerslake, Martin, & Neill, 1999) but can at the same time be quite poor indicators of performance in specific areas of mathematics (Pirie, 1987). Furthermore, a test can be valid for some students but not others. For example, a test of mathematics with a high reading demand may support valid inferences about mathematical ability for fluent readers, but when students with less developed reading skills perform poorly on the test, we cannot know whether their poor performance was due to an inability to read the items or to their weaknesses in mathematics.

Moss, Girard, and Haniford (2006), in their extensive review of the development of theories in validity, point out that the development of the concept of validity in educational assessment can be traced through successive editions of two key publications. The first originally appeared as the "Technical Recommendations for Psychological Tests and Diagnostic Techniques" in a supplement to *Psychological Bulletin* (American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement Used in Education [NCME], 1954), with new editions appearing in each subsequent decade (APA, AERA & NCME, 1966, 1974, 1985; AERA, APA, & NCME, 1999). The second is the book

Educational Measurement, first published almost 60 years ago (Lindquist, 1951) and its three subsequent editions (Thorndike, 1971; R. L. Linn, 1989; Brennan, 2006). Tracing the development of the concept of validity through these nine publications is beyond the scope of this chapter. Here, I focus specifically on the growing acceptance of the role of constructs as being at the heart of validity argument.

In the first edition of *Educational Measurement*, published by the American Council on Education, Cureton (1951) clarified that validity could not be an inherent property of an assessment, even taking into account the different possible universes of generalization. Rather, validity involves consideration of both the assessment and how it is used:

> The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third. (p. 621)

In the early days of the development of the theory of assessment in the first half of the 20th century, the mechanism for generalization was generally assumed to be from a sample of a well-defined domain to the remainder of the domain. So, for example, a test of multiplication facts would define a universe of multiplication facts, such as the 81 multiplication facts from 2×2 to 10×10, and the test would sample randomly from these 81 elements in the domain. If a student scored 50% on a test made up of 10 randomly sampled items from the domain, then the best estimate we can make is that the student knows half of the 81 multiplication facts, and furthermore, we can use the laws of statistical inference to generate confidence intervals about how accurate our estimate is likely to be.

Within such a view of validity—which Cronbach and Meehl (1955, p. 282) point out involves "acceptance of the universe of content as defining the variable to be measured"—verifying the validity of an assessment is a relatively straightforward process. It requires establishing that the items selected are *relevant* to the domain, that the collection of items included in the test is *representative* of the domain, and that enough items are included to provide an *adequate* sample.

However, this content-based approach, although being straightforward, is of very limited applicability, because most assessments are designed to assess domains far more complex than multiplication facts, so that the universe of generalization could not be defined in such a precise manner. Other approaches to the design of assessments avoided the issue of the definition of the domain entirely and instead focused on the extent to which the results of an assessment correlated with other outcomes, so that the assessment could be used to predict future performance or performance on another assessment at the same time. Thus, an assessment such as the College Board's SAT would be validated simply by the extent to which it predicted performance in higher education, for example, by examining the correlation between SAT scores and college freshman grade point average (GPA). A group-administered test of dyslexia that provided similar results to an individually administered 3-hour clinical

interview undertaken by an educational psychologist (Miles, 1998) would be validated by examining the correlation between the outcomes on the two assessments. As Bechtoldt (1951) noted, acceptance of criterion-related approaches to validity "involves the acceptance of a set of operations as an adequate definition of whatever is to be measured" (p. 1245), a view encapsulated in Guilford's (1946) observation that "in a very general sense, a test is valid for anything with which it correlates" (p. 429).

These two approaches to validation were termed *predictive validity* and *concurrent validity* and, because both approaches relied on the examination of the relationship between a predictor and a criterion variable, were often grouped together under the heading of *criterion*-related validity.

The problem with these *content*- and *criterion*-related approaches to validity is that there are many things that we might want to assess for which there is no clear definition of the domain nor is there an obvious correlate that we could use to check that the assessment is doing what it is meant to do. To address this, a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement Used in Education (which later became the National Council on Measurement in Education) proposed an alternative method of validation—termed *construct* validation—that could be used where "the tester has no definitive criterion measure of the quality with which he [sic] is concerned and must use indirect measures to validate the theory" (APA, AERA & NCME, 1954, p. 14).

The idea of construct validity was further elaborated by Cronbach and Meehl (1955), who suggested that it was involved "whenever a test is to be interpreted as a measure of some attribute which is not 'operationally defined.' The problem faced by the investigator is, 'What constructs account for the variance in test performance?'" (p. 282).

Cronbach and Meehl defined a construct as

> some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct. We expect a person at any time to possess or not possess a qualitative attribute (amnesia) or structure, or to possess some degree of a quantitative attribute (cheerfulness). A construct has certain associated meanings carried in statements of this general character: Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or a few simple propositions, used in absolute propositions or probability statements. (pp. 283–284)

As Bechtoldt (1959) points out, this definition involves (at least) three characteristics:

> First, it is a *postulated attribute* assumed to be reflected in test performances; second, it has *predictive* properties; and third, the *meaning* of a construct is given by the laws in which it occurs with the result that clarity of knowledge of the construct is a positive function of the completeness of that set of laws, termed the nomological net. (p. 623)

Given the breadth of this definition of the term *construct*, it is not surprising that two years later Jane Loevinger (1957) suggested that construct validity was in fact "the

whole of validity from a scientific point of view" (p. 636), not least because Cronbach and Meehl's definition explicitly included criterion-related aspects of validity and implicitly included content considerations as well. Although there were critiques of the notion of construct validity at the time (e.g., Bechtoldt, 1959), Angoff (1988) notes that by the late 1970s, Loevinger's view "became more generally accepted" (p. 28) and there is today a broad agreement that "construct validity is indeed the unifying concept of validity" (Messick, 1980, p. 1015).

Wiley (2001) points out that, in fact, the term *construct* has been used in the psychometric literature in two ways and for two distinct purposes. The first is "to name the psychological characteristics actually estimated by an existing test score or other measurement" and the second is "to name the psychological characteristics that a test score or other measurement is intended ('designed') to measure" (p. 212).

One consequence of this consensus about the central role of construct validity is that there is substantial agreement that construct interpretations should be at the heart of all assessments. This is made clear in the latest version of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999):

> Evolving conceptualizations of the concept of validity no longer speak of different kinds of validity but speak instead of different lines of validity evidence, all in the service of providing information relevant to a specific intended interpretation of test scores. Thus many lines of evidence can contribute to an understanding of the construct meaning of test scores. (p. 5)

More recently, Wiley (2001) defined a construct as follows:

> A construct, here, is an ability (i.e., is a human characteristic required for successful task performance). At the simplest level, these constructs can be identified with capacities to perform classes of tasks defined by task specifications. Because they must enable more than a single task performance, the concept implicitly follows from the formulation of an equivalence class of task implementations or realizations, all of which require possession of the same ability construct for successful performance. However, in order to be an ability, a human characteristic must not only differentiate successful from unsuccessful task performance, but must also apply to some tasks and not to others. That is, every ability must be defined so that it subdivides tasks into two groups: those to which that ability applies and those to which it does not. (p. 208)

Assessment is contentious, therefore, because when we assess we go beyond the construct and claim that certain tasks, if performed successfully, indicate the presence of the ability in the individual, although if the individual does not perform the task successfully, this is taken as evidence that the individual does not have the ability in question. In other words, *assessments operationalize constructs*. It is often the case that there is broad agreement on the matter of curriculum, although there is considerable disagreement about the associated assessment because it removes the ambiguity about the construct being assessed. These disagreements frequently manifest themselves as debates about assessments. However, the argument of this chapter is that these disagreements are in general better thought of as differences in what should be the construct of interest, and it is solely because the construct is made manifest only in the assessment does it appear that the validity of the assessment is a

matter of opinion. This idea can be illustrated by looking at the issue of gender bias in the assessment of history in secondary school.

AN INTRODUCTORY EXAMPLE: THE ASSESSMENT OF HISTORY

Breland (1991) found that males outperformed females when achievement in history was assessed with multiple-choice tests but that females outperformed males when achievement in history was assessed with constructed-response items. One (quite common) interpretation of this is that multiple-choice tests are biased against female students, and the differential performance of males and female students is therefore a question of the validity of the assessments used. However, such debates can also be examined as debates over construct definition; in other words, when we assess history, what, in fact, are we really assessing? One view is that achievement in history is primarily about "facts and dates." For adherents to this view (see, e.g., McGovern, 1994), multiple-choice tests are highly appropriate ways to measure achievement in history, because they allow a wide range of knowledge to be assessed quickly and efficiently, and they have the additional advantage that the scoring is objective (apart from issues related to erasures, ambiguous marks, etc.). Another view is that history is primarily about the integration of partial (in both senses: partisan and incomplete) and sometimes conflicting sources of evidence to assemble a descriptive, and ideally explanatory, account of historical events (Wineburg & Fournier, 1994). For adherents to this second view, multiple-choice tests are likely to be inadequate as measures of achievement in history because it is extremely difficult, if not impossible, to assess such thinking through multiple-choice tests. Debates between adherents to these differing views of history often manifest themselves as apparently technical discussions about the validity of particular approaches to assessment, but the argument of this chapter is that they are more productively viewed as arguments about construct definition. Like this debate about history, many debates about the adequacy and appropriateness of assessments appear, on the surface, to be debates about technical issues, but they are, in fact, debates about construct definition. The reason that this observation is important is that debates about construct definition cannot be resolved by those with expertise only in assessment. They are debates outside assessment that should be settled before the assessment is designed. Otherwise, what is easy to assess, what is practicable to assess, and what is inexpensive to assess will be unduly influential in the determination of the assessment. This is not to say that these questions are unimportant; only that they must not be allowed to predominate, or prejudice, discussion of other aspects of the quality of assessment. The case of the assessment of history discussed above can be used to illustrate the two main threats to the validity of construct interpretations of assessment outcomes: *construct underrepresentation* and *construct-irrelevant variance* (Messick, 1989)—in other words, whether the assessment is too narrow to support the intended construct interpretations or whether the assessment systematically introduces extraneous information into the scores (McCallin, 2006).

Proponents of the view of history as "facts and dates" may well view multiple-choice assessments as entirely appropriate because they are able to assess all of what they believe history to be through such tests. Proponents of the view of history as "interpreting evidence" will, however, regard such tests as underrepresenting the construct of history; in other words, they argue that inferences one might make about a student's knowledge of facts and dates on the basis of a multiple-choice test may well be valid, but inferences about the ability to assemble historical arguments are much less likely to be warranted, because important aspects of the construct of interest were not assessed. Adherents to the "interpreting history" view would regard that multiple-choice test as suffering from *construct underrepresentation*.

Conversely, proponents of the view of history as "facts and dates" are likely to view assessments that involve items that require extended constructed responses as equally flawed. This is because although such assessments may assess aspects of the construct of interest, they are also assessing other capabilities, such as the ability to convey meaning in writing. All assessment users want the differences in students' scores to reflect differences in the construct of interest, but if the differences in scores are also, in part, attributable to factors unrelated to the construct of interest, then construct interpretations of students' scores are problematic. Advocates of the "facts and dates" view of history observe that although variation in students' outcomes on constructed-response tests of history do, in part, measure knowledge of history, differences in scores also represent differences in the ability of students to write well and even, perhaps, skill in handwriting. In other words, the students' scores would suffer from *construct-irrelevant variance*.

So while the "interpreting evidence" lobby would regard scores on multiple-choice tests as underrepresenting the construct of interest, the "facts and dates" lobby would regard scores on extended constructed response assessments as introducing a degree of construct-irrelevant variance.

Returning to the issue of gender bias, it can be seen from the above argument that the original question about whether multiple-choice tests of history are biased against females resolves into two separate issues, which need to be addressed in different ways. The first is, "What is the construct of history to be assessed?" This is not a technical issue. It is essentially a philosophical issue about the nature of history as an academic discipline (and therefore a specific version of the question, "What is knowledge?"). This is a matter on which individuals can legitimately disagree. The second question is, once a particular view of the construct of history has been agreed, whether a particular set of assessment arrangements adequately addresses the construct.

The advantage of such a formulation is that it clarifies the nature and the origin of any gender difference. If particular assessments show markedly superior performance for boys than for girls, is this because the assessment arrangements have introduced a degree of construct-irrelevant variance into the scores, for example, by choosing historical topics that are of greater interest to boys than girls? Or was the construct of interest defined in such a way that it is intrinsically something at which boys are better than girls?

Similar issues arise in the case of the mental rotation of three-dimensional objects. The ability to determine whether two three-dimensional objects are the same shape but oriented differently, or whether they are different objects, has been extensively studied, and this ability shows marked sex differences, with males outperforming females (often by as much as one standard deviation) in almost all cultures (Voyer, Voyer, & Bryden, 1995). This has led some to suggest that tests of this ability are biased against females. But this is to locate the problem in the wrong place. A test tests only what a test tests, and in this example, males really are better than females at this particular skill. The bias is not in the test but might be present in the inferences based on its outcomes. So whether such items should be included in a mathematics test is an issue of construct definition. The inclusion of items on mental rotation of three-dimensional objects will, for some, introduce a degree of construct irrelevant variance, whereas for others, the exclusion of such items would introduce construct underrepresentation. The important point is that the debate should be focused on the issue of construct definition, and the consequences of the definition, rather than on the technical issues of the extent of construct-irrelevant variance and construct underrepresentation. The ability to rotate three-dimensional objects mentally clearly seems related to mathematics, so the definition of the construct of mathematics in such a way as to include this topic would appear to be justified in terms of an appropriate philosophical imperative for the discipline of mathematics. On the other hand, including in the definition aspects of mathematics at which males are known to be better than females, with a corresponding set of messages about who can and cannot succeed in mathematics, would appear to run counter to a moral imperative about equity. Assessments of mathematics that include mental rotation are not biased, because it ascribes to an assessment a property it cannot have—what Ryle (1949) called a "category error."

There is one more important aspect of the relationship between constructs and the assessments that are intended to assess them, and that is that these relationships can change over time. One common response to the suggestion that a particular assessment suffers from construct underrepresentation is to point out that although the assessment itself might not adequately represent the whole of the construct of interest, the items actually assessed do in fact function as an adequate proxy for those aspects of the construct not assessed (in essence an argument for concurrent validity). For example, the "facts and dates" community might retort that an individual's performance on the multiple-choice test turns out to be quite a good proxy for performance on other aspects of history, such as those assessed with constructed-response tasks. However, the fact that such correlations are observed does not mean that such correlations are likely to be the same in the future. For example, in England, the most commonly used measure of school performance is the proportion of students from that school achieving one of the four highest (of nine) grades in at least five school subjects in the national school leaving examination (the General Certificate of Secondary Education or GCSE). For many years, the percentage of students achieving this benchmark including passes in English and Mathematics was around 10 percentage

points lower than if the benchmark included any five school subjects. However, from 2001, as pressure to increase results from government increased, schools increasingly looked for ways to increase student success by matching the choice of subjects taken to the specific aptitudes of the students. By 2008, the effect of this was to increase the "gap" between the two indices (i.e., success rates including, and not including, passes in English and Mathematics) from 10 percentage points to 16 (Department for Children, Schools and Families, 2008). In other words, the relationship between the two measures of success and, in particular, the ability to predict one from the other had changed as a result of the social processes in play. This is, of course, just another example of Campbell's (1976) law:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (p. 49)

The fact that the relationship between assessment outcomes can change over time was one of the reasons that Messick (1989) introduced the idea that the validity argument should include a consideration of the consequences of result interpretation and use, although this has been widely misunderstood and misapplied. Messick was not suggesting that all social consequences of test interpretation and use should be considered part of validity, but only those that were attributable to weaknesses in the test.

> As has been stressed several times already, it is not that adverse social consequences of test use render the use invalid, but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct-irrelevant variance. If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced—or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible—then the validity of the test use is not overturned. Adverse social consequences associated with valid test interpretation and use may implicate the attributes validly assessed, to be sure, as they function under the existing social conditions of the applied setting, but they are not in themselves indicative of invalidity. (pp. 88–89)

To see how this formulation of validity works in practice, let us return to the example of history in secondary schools discussed above. Suppose that a state defines the construct of competence at a particular grade in history much along the lines of those advocated by proponents of the "interpreting evidence" view of history, but finds that the costs of constructed response tests to assess this are too great, and therefore decide to rely on multiple-choice tests. The state then defends this course of action on the grounds that concurrent validity studies have shown that scores of students on constructed-response assessments of history are highly correlated with those on the multiple-choice tests. If teachers in the state subsequently change the way that they teach history and focus on the "facts and dates" that are the focus of the test, then the test will support less adequately inferences about aspects of history related to "interpreting evidence," and it may also be the case that female students enjoy history less because the subject is less connected to their existing knowledge

(Belenky, Clinchy, Goldberger, & Tarule, 1986; Gilligan, 1982). In this example, the undesirable consequences were at least in part caused by the fact that the test in use underrepresented the construct of history defined by the state, and so the validity of the state-mandated test is in question. On the other hand, had the state *defined* history as facts and dates, then the fact that such a definition—even though validly assessed with multiple-choice tests—resulted in the marginalization of female students would not compromise the validity of the assessment. It would, however, bring into question the ethical defensibility of defining history in a way that was likely to exclude female students. This might in turn generate debate related to "objectivist" versus "constructivist" views of knowledge (von Glasersfeld, 1995), but it is the thesis of this chapter that it would be more productive to have the debate in these terms rather than in terms of the assessment. Of course, where knowledge is defined as socially constructed, rather than being objectively defined, then construct definition may become more difficult—or at least less straightforward—but unless assessment design begins with construct definition, then establishing the inferences that may validly be drawn from the assessment outcomes becomes difficult, if not impossible.

In the remainder of this chapter, I explore the interplay of constructs and assessment in three broad areas of current interest in assessment. These specific areas have been chosen because they are areas of significant current debate, are areas where there is significant potential for inequity, and, perhaps unsurprisingly, are areas where the particular focus of this chapter provides a useful perspective. Another potential area for such an analysis—that of the testing of English language learners—has been the subject of a thorough review in a previous issue of *Review of Research in Education* (Durán, 2008). As noted in the introduction to this chapter, the three areas are

- Testing for admission to higher education, focusing specifically on impact on people of color
- The rise, and fall, of measures involving constructed-response items (including portfolio assessment, authentic assessment, and performance assessment) in large-scale achievement testing within statewide accountability systems
- The issue of accommodations in mandated assessment for students with special educational needs, especially within the context of the No Child Left Behind Act

Within each of these areas, the chapter explores the interplay of issues of technical adequacy with equity and suggests that a focus on the consequences of particular choices of constructs, rather than on the assessments, will lead to more helpful descriptions of what is happening and what can be done to improve equity in education.

EQUITY IN HIGHER EDUCATION ADMISSIONS TESTING

Recruiting and selecting students for higher education is a global challenge (Kellaghan, 1996). In many—and perhaps most—countries, this is achieved through

the provision of a set of examinations of academic achievement that are either set by higher education institutions (HEIs) or over which HEIs exert a considerable influence (Eckstein & Noah, 1993). Providing a set of standardized procedures for assessing the suitability of students for admission to university has had the effect, in most countries, of establishing a common curriculum for the period of upper secondary schooling (typically the last 2 or 3 years of secondary school). Although such systems may allow for different pathways (e.g., Sweden), constrained choice (e.g., Germany), or even unrestricted choice of subjects (e.g., England), such systems typically share a common "tariff" that allows the results of different students to be placed—more or less meaningfully—on a single scale.

A major factor in the evolution of these assessment systems was that in these countries upper secondary education had been designed from the outset primarily as a preparation for higher education and therefore intended only for the educational "elites" (for most of the 20th century less than 10% of the population). In the United States, however, between 1910 and 1940, there was a massive expansion in the provision of upper secondary education for all students, as a preparation for citizenship. It is notable that in many states in the United States the rate of participation of 18 year olds in education achieved in the 1930s has still not been reached in many rich European countries (Goldin & Katz, 2008; Organisation for Economic Cooperation and Development, 2008).

The fact that upper secondary education was, in the United States, primarily a local matter created few difficulties during much of the 19th century for three main reasons. First, many universities appeared to be recruiting rather than selecting students, so the criteria for entry were as much financial as academic (Levine, 1986). Second, students tended to apply to universities in the same state as the high school they had attended, which made it possible for the universities to assure the quality of entrants, either by accrediting the high schools and their grading standards (as piloted by the University of Michigan) or to set a statewide university entrance examination, such as those instituted by the Board of Regents of the State of New York in 1878. Third, those universities that did draw significant numbers of students from outside the state had their own entrance examinations; Harvard and Yale, for example, had introduced entrance examinations in 1851 (Broome, 1903).

As applications for university places grew in number, and the pattern of applications increased in complexity, a group of eight elite universities now known as the "Ivy League" (Brown, Columbia, Cornell, Dartmouth, Harvard, Pennsylvania, Princeton, and Yale) proposed the establishment of a set of common entrance examinations and established the College Entrance Examinations Board to take this work forward.

For the first four decades of its existence, the College Board relied primarily on written examinations of scholastic achievement based on the traditional school subjects. However, in 1934, Harvard had started using the Scholastic Aptitude Test (SAT), which had been developed by Carl Campbell Brigham from the intelligence tests administered to army recruits during the First World War (Zenderland, 1998),

for its national scholarship awards. The early success of the SAT in this context—students who had scored highly on the SAT did well at Harvard—led, in 1937, to the adoption of the SAT for all scholarship decisions at 14 of the universities that were then members of the College Board (Hubin, 1988). In 1941, the "College Boards"—the traditional written examinations that had been in use since 1901—were withdrawn, leaving the SAT as the dominant university entrance test in the United States. A detailed discussion of the development of the SAT is beyond the scope of this chapter. A good account of the early development can be found in Hubin (1988), and Lemann (1999) provides details of the more recent history. Here, I focus, in particular, on issues of construct definition and equity.

The SAT is probably "the most researched test in the world" (College Board, 2009). A review of more than 3,000 studies of the validity of the SAT as a predictor of performance in the early years of college (Hezlett et al., 2001) found coefficients ranging from .44 to .62 and also predicted a range of measures of performance later in college such as likelihood of graduation and cumulative GPA, although the values of the coefficients were lower (ranging from approximately mid-30s to mid-40s). Although high school grade point average (HSGPA) appears to be the best single predictor of a student's GPA in their first year at college, using the SAT in addition does improve the prediction. A study of around 20,000 students who took the SAT in 1995 (Kobrin, Camara, & Milewski, 2002) found values of the correlation coefficients, corrected for attenuation of range (Thorndike, 1949) and shrinkage (Vogt, 1999), ranging from .30 for Hispanic students to .41 for Asian American students for the HSGPA and from .31 for Hispanic students to .44 for Asian American students for the SAT. The combination of HSGPA and SAT, however, does considerably better, ranging from .42 for Hispanic students to .55 for Asian American students (.48 for White and .50 for African American students).

The SAT has been widely attacked on a number of grounds. Some criticisms point out that although the SAT does increase the accuracy of prediction, the increases are marginal. For example, Crouse and Trusheim (1988) used a sample of 2,470 students from the National Longitudinal Study (NLS) of the class of 1972 to examine the utility of the SAT in predicting which students will earn a first-year college GPA of 2.5 or better. They found that forecasts made without the use of high school rank were accurate in 53% of the cases, and those made with high school rank had an accuracy of 62.2%, an increase of 9.2 percentage points. The use of the SAT in addition to high school class rank improves the accuracy by a further 2.7 percentage points. Although this is a proportional increase of almost 30%, Crouse and Trusheim suggest that such a small absolute increase in the accuracy of the prediction would not be noticed by most colleges. They also suggest that achievement tests, rather than tests of so-called aptitude, would have more beneficial social consequences, although the correlation between the SAT and the best-known national achievement test—the ACT—at .92 is comparable to the test–retest reliability of the SAT itself (Dorans, 1999).

Other concerns focus on the fact that scores for most minority students are lower—and in the case of some students, much lower—than for White students,

with differences of as much as one standard deviation. The study by Kobrin et al. (2002), discussed above, found that African American and Hispanic students scored, respectively, 1.19 and 0.98 standard deviations below White students. Perhaps the best known of recent attacks has been that by Freedle (2003), wherein he states that "the SAT has been shown to be both culturally and statistically biased against African Americans, Hispanic Americans, and Asian Americans" (p. 1). This statement is inaccurate and misleading in several ways. First, the statistical arguments made by Freedle have been shown to contain important errors. Freedle's analysis focuses on specific types of items and attempts to show that they are more difficult for educationally disadvantaged students and proposes a way of rescoring the SAT (essentially placing greater weight on more difficult items) that raises the score of many minority students (by more than two standard deviations for some minority students on the verbal section of the SAT). Dorans (2004) points out several problems in Freedle's analysis. First, the version of the SAT analyzed by Freedle dated from 1980, before the routine screening of items for "differential item functioning" (Holland & Wainer, 1993)—the idea that an item might function differently for different groups of students of the same overall proficiency—was introduced.

The most significant problem, however, is that for each item a different group of students is used in comparing the success rate of Black and White students, and "the groups used to study a question vary from question to question in a systematic way related to the difficulty of the question and the proficiency of the examinees answering the question" (Dorans, 2004, p. 65). When the items are compared on the same basis, the effects found by Freedle disappear almost entirely (the remaining effect is of the order of one twentieth of a standard deviation). The cause of the small differences that remain are not fully understood but are possibly related to the fact that Black students are less likely to attempt the harder items on the test and are therefore less likely to be penalized for incorrect responses.

Second, and more important in terms of the arguments presented in this chapter, the statement by Freedle is a form of category error, because bias is not a property of assessments but of the inferences that are made on the basis of their outcomes. Even if Freedle's argument is reframed in terms of inferences rather than a statement about the SAT itself, it is incorrect, because the SAT actually *over*predicts performance at college for most minority students on average (Kobrin et al., 2002). This is where the importance of adequate construct definition is most clearly demonstrated. Minority students do less well on the SAT because they do less well in college, and the SAT is designed to predict performance in college. The SAT does underpredict college performance for females, but so does HSGPA (Kobrin et al., 2003). The extent to which HSGPA, SAT, and a composite of the two overpredict and underpredict first-year college GPA at 23 higher education institutions for males and females of different minorities is shown in Table 1 (based on data from Kobrin et al., 2003).

Perhaps the most remarkable feature of the data in Table 1 is that for every group apart from "Other" the SAT either overpredicts or provides a more accurate prediction of first-year college GPA. The broad picture is that the SAT does improve the

TABLE 1
Overprediction and Underprediction of First-Year College GPA in 23 Institutions

	HSGPA		SAT		HSGPA + SAT	
	Female	Male	Female	Male	Female	Male
African American	.09	.20	.01	.22	−.01	.16
Native American	.13	.24	.06	.32	.07	.28
Asian American	.03	.08	−.01	.15	−.03	.07
Hispanic	.23	.31	.03	.20	.04	.20
White	−.11	−.03	−.11	.06	−.09	.05
Other	−.09	.04	−.13	.03	−.12	.04

Source. Derived from Kobrin et al. (2003).
Note. GPA = grade point average; HSGPA = high school grade point average; SAT = Scholastic Aptitude Test.

accuracy of prediction of performance in the first-year performance at college when compared with HSGPA alone (although not by very much).

The central thesis of this chapter is that many debates in education present themselves as debates about the validity of assessments but are more productively conceptualized as issues of construct definition and construct choice. The SAT is not biased against, for example, African American students not only because bias is not a property of tests but also because for both males and females the SAT actually overpredicts college performance. The purpose of this argument is not to defend the SAT but to frame the debate in a way that is more likely to lead to productive action.

If we read the evidence about the differential impact about the SAT on minorities as being caused by the SAT, then that will focus our attention on measures such as those proposed by Freedle, aimed at reducing Black–White score differences. Although increasing the degree of randomness in a selection procedure may serve to reduce adverse impact on minorities, selecting students for programs on which they are likely to be unsuccessful would seem to do little to advance the cause of equity (Wiliam, 2003).

The argument of this chapter is that a more productive way of reading the evidence from the predictive validity studies of the SAT is in terms of the construct being assessed. In other words, the main reason that many minority groups do less well on the SAT (as well as having lower HSGPA) is because they are less prepared for college, and indeed, scores on the National Assessment of Educational Progress show differences of approximately the same magnitude between White students, on one hand, and African American or Hispanic students on the other, as are found on the SAT—approximately one standard deviation (although the gap between White and minority students has been closing). Such a large gap is not surprising given the fact that minority students tend, on average, to be resourced less advantageously than White students (Piché & Taylor, 1991), are less likely to be taught by well-qualified

teachers (Ferguson, 1991), and are more likely to live in conditions that affect cognitive development (Hackman & Farah, 2009; Kishiyama, Boyce, Knight, Jimenez, & Perry, 2009).

However, as David Hume (1973/1896) pointed out in his *Treatise on Human Nature*, one cannot deduce an "ought" from an "is." The fact that the SAT currently does predict reasonably who will, and who will not, do well at college does not mean that we should accept this uncritically. Clearly, it *is* currently the case that levels of school achievement, whether measured by HSGPA or the SAT, predict college achievement well, but this is not in itself evidence of equity when the opportunities for high-quality instruction and support are themselves not equitably distributed. No test is perfect, and indeed, it could be argued that the SAT has more than its share of faults, but blaming the SAT for the failure of students of color to gain admission to, and to thrive in, the most selective colleges is unlikely to do anything to improve the situation. The SAT works as well as it does because it is exquisitely tuned to the system in which it operates. Lowering admissions requirements for college while leaving the educational systems in the college unchanged is likely to do little except increase the number of students failing to complete their studies. Instead, because this is a system problem, systemic approaches are likely to be more successful (O'Day & Smith, 1993). An example of such a systemic approach is provided by the *Access to Medicine* program offered at King's College London (Access to Medicine, 2009).

The medical program at King's College London (part of the University of London) is one of the largest in Europe, graduating approximately 350 doctors every year. The campuses of the College are located in some of the most ethnically diverse parts of London, but prior to the introduction of the Access to Medicine program very few students attending local schools were admitted, and those that were admitted were more likely to be of Asian descent rather than the most common minorities in the area—those of African heritage. As well as raising the issue of equity, such selection policies were unlikely to result in culturally competent health services (Council of Heads of Medical Schools, 1998).

To address this, the Access to Medicine program addressed simultaneously the three issues of *recruitment*, *selection*, and *retention* (Wiliam, Millar, & Bartholomew, 2004). Recruitment was addressed through intervention programs implemented in local schools with students at the beginning of secondary school (age 11) to raise aspirations and to ensure that curricular choices made by students at the ages of 14 and 16 did not "close off" particular routes into medical education. This was particularly important because informal contacts with schools had suggested that many students aspiring to be doctors thought that the most important subject to study in upper secondary school was biology, whereas, in fact, the only subject required for admission to a medical program by the General Medical Council of the United Kingdom is chemistry.

At the time of the inception of the Access to Medicine program, selection to most medical schools in the United Kingdom was based principally on the grades achieved by students at the Advanced level of the General Certificate of Education examination

(usually abbreviated to "A-level"). Students in the local schools rarely achieved the grades required for admission, but lowering the entry requirements for students from local schools would be likely to admit students who were not well prepared, or indeed well suited, to the intense study required in the medical degree program at King's College London.

Because a great deal of the curriculum for the first 2 years of the medical program at King's focuses on basic medical sciences, it was felt that assessing the ability to reason scientifically and specifically the ability to integrate new ideas into existing scientific schema might provide a way of identifying aptitude for medical education even where students had not been well taught. A number of science reasoning tasks, developed by Shayer and Adey (1981), had proved to be very successful at predicting later science learning, and so these tests were used to identify students from local schools who did not meet the traditional criteria for admission to a medical degree program but who did show the ability to learn science quickly. To counter charges of "dumbing down," cut-scores on the science reasoning tasks were established by reference to the cohort of medical students admitted on the traditional basis.

Although the students selected on such a basis were likely to have scientific reasoning skills on par with traditional students, they were much less likely to have a secure grounding in basic scientific knowledge, particularly in chemistry. The students selected for the Access to Medicine program were also less well prepared for higher education in general. To address this it was decided that the Access to Medicine students would be allowed 3 years to cover the same curriculum followed by traditional students in the first 2 years of the 5-year medical degree program. Philanthropic sources of funding were secured to pay for the cost of the extra year (both tuition and living expenses), and the students were given additional support in the form of a part-time tutor.

The first students admitted under the Access to Medicine program graduated in 2007. Although the early cohorts are too small to generate any robust findings, initial outcomes are encouraging: the final scores of the Access to Medicine students are indistinguishable from those of the "traditional students" (Garlick & Brown, 2008).

For the purpose of this chapter, the most important feature of the Access to Medicine program is that instead of focusing on the assessment instrument used in selecting students for admission to medical programs (the A-level), the program used a systemic approach, involving recruitment, selection, and retention. Rather than blaming the instrument for the low scores achieved by minority students, the A-level was accepted as an adequate basis for recruiting and selecting students who had been well prepared for higher education and thus would do well in the traditional medical program. To increase equity in the outcomes for minority students, the program looked for different constructs that might be used in predicting success for nonstandard students, in the context of a nonstandard program.

In the same way, it seems likely that the advancement of underrepresented minorities in American higher education will be better secured by a search for new ways of raising aspirations for students and through new kinds of higher education curricula,

supported by new forms of assessment, thus acknowledging the diversity of students' previous experiences, rather than further efforts attempting to show that one particular instrument is biased against minority students.

THE RISE AND FALL OF AUTHENTIC ASSESSMENT IN THE UNITED STATES

Although the provision of education has always been regarded as an essentially "local" matter in the United States, over the last 50 years, state and federal sources have become greater and greater net contributors (Corbett & Wilson, 1991). Despite the fact that the annual polls conducted by the *Phi Delta Kappan* organization have indicated that most parents are happy with their local schools, many states felt the need to make school districts accountable beyond the local community, through the introduction of statewide testing programs. For example, in 1961 California introduced a program of achievement testing in all its schools. Although the nature of the tests was at first left to the districts, in 1972, the California Assessment Program was introduced, mandating multiple-choice tests in English language, arts, and mathematics in Grades 2, 3, 6, and 12 (tests for Grade 8 were added in 1983). Subsequent legislation in 1991, 1994, and 1995 enacted new statewide testing initiatives that were only partly implemented. However, in 1997, new legal requirements for curriculum standards were passed, which, in 1998, led to the Standardized Testing and Reporting program. Under this program, all students in Grades 2 to 11 take the Stanford Achievement Test—a battery of standardized tests—every year. Those in Grades 2 to 8 are tested in reading, writing, spelling, and mathematics, and those in Grades 9, 10, and 11 are tested in reading, writing, mathematics, science, and social studies. In 1999, further legislation introduced the Academic Performance Index—a weighted index of scores on the Stanford Achievement Tests, with awards for high-performing schools, and a combination of sanctions and additional resources for schools with poor performance. The same legislation also introduced requirements for passing scores on the tests for entry into high school and for the award of a high school diploma.

There were some legal challenges to the notion of "minimum competency" testing, most notably in Florida, where in 1978, a student, named Debra P, brought a case against the state commissioner of education, Ralph Turlington, and others because she had been denied a high school diploma on the grounds that she had failed to pass a minimum-competency test required by the state (United States District Court 474 F. Supp. 244 M.D. FL, 1979). The key point in the case was that Debra P was Black, and when she began her education in 1967 she had attended a segregated elementary school, which had been resourced less favorably than the schools attended by Whites. In its final judgment, the court decided that the requirement to pass a minimum-competency test placed a greater burden on a Black student than a White student and was therefore unfair. The court decided that the State of Florida could not deny students high school diplomas for another 4 years from the date of the judgment, by

which time the court believed all students would have had adequate opportunity to learn the material on which the test was based. Provided states were prepared to be able to show that all students did have the opportunity to learn the material covered in the tests, minimum-competency requirements for high school diplomas were fair.

At the same time, many states were experimenting with alternatives to standardized tests for monitoring the quality of education and for attesting to the achievements of individual students. In 1974, the National Writing Project (NWP) had been established at the University of California, Berkeley (Lieberman & Wood, 2002). Drawing inspiration from the practices of professional writers, NWP emphasized the importance of repeated redrafting in the writing process, and to assess the writing process properly, one needed to see the development of the final piece through several drafts. In judging the quality of the work, the degree of improvement across the drafts was as important as the quality of the final draft.

The emphasis on the process by which a piece of work was created, rather than the resulting product, was also a key feature of the Arts-PROPEL project—a collaboration between the Project Zero research group at Harvard University (Gardner, 1989) and Educational Testing Service. The idea was that students would "write poems, compose their own songs, paint portraits, and tackle other 'real-life' projects as the starting point for exploring the works of practicing artists" (Project Zero, 2005). Originally, it appears that the interest in portfolios was intended to be primarily formative, but many writers also called for performance or authentic assessments to be used instead of standardized tests (Berlak, 1992; Gardner, 1992).

Two states in particular, Vermont and Kentucky, did explore whether portfolios could be used in place of standardized tests to provide evidence for accountability purposes, and some states also developed systems in which portfolios were used for summative assessments of individual students. However, the use of portfolios was attacked on several grounds. Chester Finn, President of the Thomas B. Fordham Foundation, said that portfolio assessment was "costly [. . .] slow and cumbersome" and went on to say "its biggest flaw as an external assessment is its subjectivity and unreliability" (Mathews, 2004, p. 75).

In 1994, the RAND Corporation released a report on the use of portfolios in Vermont (Koretz, Stecher, Klein, McCaffrey, & Deibert, 1994), which is regarded by many as a turning point in the use of portfolios (Mathews, 2004). Koretz et al. (1994) found that the meanings of grades or scores on portfolios were rarely comparable from school to school because there was little agreement about what sorts of elements should be included. During the short time the portfolios had been in use, the standards for reliability that had been set by standardized tests such as the SAT simply could not be matched. Although advocates might claim that portfolios were more valid measures of learning, the fact that the same portfolio would get different scores according to who did the scoring made their use for summative purposes difficult to sustain in the U.S. context.

In fact, even if portfolios had been able to attain high levels of reliability, it is doubtful that they would have gained acceptance. Teachers did feel that the use of

portfolios was valuable, although the time needed to produce worthwhile portfolios detracted from other priorities. Mathematics teachers in particular complained that "the mathematics portfolios required a significant amount of class time, which had to be taken from other activities" (Koretz et al., 1994, p. 26). Furthermore, even before the RAND report, the portfolio movement was being eclipsed by the push for "standards-based" education and assessment (Mathews, 2004).

For the purposes of this chapter, one of the most interesting features of the decline of performance assessment is that construct considerations played a much smaller part in the discussion than technical considerations (such as reliability) or issues of manageability. Even on narrow technical considerations, the widespread rejection of portfolios for high stakes accountability testing in the United States would appear to have been premature, for two reasons.

First, the experiences of many other countries are that, with appropriate sources of support, and given sufficient time, portfolio assessment can attain similar levels of reliability to those achieved in standardized tests. One unpublished investigation into the assessment of portfolios of work in English language arts for 16-year-old students in England found levels of classification accuracy on a 9-point scale of approximately 70% (exact match). Although this was considered rather unsatisfactory (and may have contributed to the decision not to publish), it corresponds to a classical reliability of approximately .92 (Wiliam, 2001)—higher than is achieved on many—if not most—accountability tests.

Second, the experience of other countries is that, given time, teachers are also able to integrate the portfolio work into their ongoing instruction, so that, instead of being seen as an unmanageable addition to the curriculum, portfolios come to be seen as a vehicle for delivering that curriculum as well as a valuable focus for teacher professional development (Maxwell, 2004).

The rejection of portfolios and other forms of assessment requiring students to construct, rather than select, responses for high stakes accountability assessment appears to have profound consequences for overall levels of educational achievement in the United States. E-portfolios are changing from being simply digital repositories to being learning environments that support a range of pedagogical practices and afford novel kinds of collaborative learning (Abrami & Barrett, 2005), which are likely to be increasingly important in the future (Jewitt, 2006).

Moreover, as Newmann, Bryk, and Nagaoka (2001) have shown, accountability systems that rely on what they term *authentic intellectual work* assessed through constructed response items are associated with higher levels of student achievement. Comparative studies suggest that whereas factors such as teacher quality appear to be the most important predictors of high scores in international comparisons of student achievement (Barber & Mourshed, 2007), the use of accountability measures based on constructed-response items measuring higher order knowledge appears to be an important additional factor (Bishop, 2001a, 2001b), although in this context it is worth noting that poorly designed performance measures, combined with very high stakes for teachers and schools, can result in undesirable outcomes (Wiliam, under review).

It would be naive to assume that a focus on constructs, rather than assessments, would have led to the retention of portfolios and other forms of authentic assessment in high stakes accountability testing. However, it seems plausible that an increased focus on what the assessments were supposed to be measuring, rather than on the assessments themselves, would have contributed to the debate and might have made stakeholders more willing to consider alternatives to standardized multiple-choice testing.

In the remainder of this section, I review briefly how issues of construct definition have interacted with equity in the case of the learning of mathematics.

For many years, the fact that the performance of males in mathematics was superior to that of girls attracted little attention, as if somehow this was the natural order of things. Maccoby and Jacklin (1974), in their monumental work *The Psychology of Sex Differences*, reviewed more than a thousand research studies and concluded that it was "fairly well established" that boys outperformed girls in terms of visuospatial ability and mathematics, whereas girls had more developed verbal abilities (pp. 351–352). Although many authors challenged both the methodology and the conclusions (see, e.g., Block, 1976), there can be little doubt that the work served as a significant impetus for further work in this field.

Although Maccoby and Jacklin (1974) made clear that their view was that these differences were caused by a range of factors, including biological predispositions, social shaping, and cognitive self-actualization processes (O'Connell, 1990), much of the subsequent work on sex differences, particularly in mathematics achievement, focused on genetic factors, often with speculations about how the environment of evolutionary adaptedness (Bowlby, 1969) might have contributed to the development of more advanced spatial skills in males. However, in recent years, the extraordinary decline in the size of sex differences in mathematics performance has provided strong evidence that observed sex differences are primarily of environmental, rather than genetic, origin.

Feingold (1988) examined the performance of males and females on the SAT from 1947 to 1983, and although there was evidence of female superiority in language and male superiority in mathematics, the magnitude of the difference declined markedly over the period studied. In fact, current estimates suggest that the magnitude of sex differences has halved over the second half of the 20th century (Hyde, Fennema, & Lamon, 1990; M. C. Linn, 1992), and a meta-analysis of 98 studies from 1974 to 1987 on sex differences in mathematics concluded that a 95% confidence interval for the effect size included zero (Friedman, 1989). Most recently, Hyde, Lindberg, Linn, Ellis, and Williams (2008) found that the mean sex difference on state assessment in mathematics across 10 states in the United States (California, Connecticut, Indiana, Kentucky, Minnesota, Missouri, New Jersey, New Mexico, West Virginia, and Wyoming) was .0065 (not significantly different from zero). There is also increasing evidence that male superiority in mathematics is most marked in countries with the greatest gender inequality (Hyde & Mertz, 2009).

These analyses of the magnitude of sex difference have tended to treat measures of mathematics achievement as interchangeable. A complementary strand of research

studies, arguably begun by Carol Gilligan's (1982) landmark book *In a Different Voice* has sought to look at how school subjects are defined and the extent to which they are resonant with the modes of thinking that are preferred by different individuals. In particular, work by Boaler (1997) has shown that while both males and females prefer to make connections between their existing and new knowledge in mathematics, doing so appears to be more important for females. Where mathematics is defined as a series of disembodied facts with no relationship to the world outside the mathematics classroom, then making such connections is difficult and available to only a few students. Where, however, mathematics is presented as being a way of thinking about real issues in a disciplined way, then mathematical thinking is available to all (Boaler, 2008).

This is, of course, an issue of construct definition, analogous to the example of history discussed previously. Absent adequate definitions of the construct of school mathematics, all assessments can be regarded as measuring the same construct. Performance assessments come to be seen as expensive, unreliable, and time-consuming measures of student achievement. It is hardly surprising, therefore, that their use has been radically curtailed in all the statewide assessments mandated by the No Child Left Behind Act of 2001 in the United States, to be replaced by multiple-choice tests.

The argument of this chapter is that had the focus been first on the construct to be assessed, and only second on the technical adequacy of the assessments developed to assess the construct, then the construct of mathematics being assessed in schools in the United States might be very different. A focus on mathematics as inquiry would produce more equal outcomes between males and females (Willingham & Cole, 1997) and between different minorities (Boaler, 2008)—and, as Burton (2004) has shown, closer to the conception of mathematics held by professional mathematicians.

ACCOMMODATIONS FOR SPECIAL POPULATIONS

As Koretz (2008) notes, "Few issues in measurement raise such intense emotions as the assessment of students with special needs: those with disabilities or those with limited proficiency in English" (p. 281). In this chapter, I will not deal with the difficulties of identification and classification of disabilities and special needs nor with general issues of assessments for such students; an excellent summary of the current "state of the art" in this area can be found in Pullin (2008). Here, in keeping with the rest of the chapter, I focus on the importance of construct definition in the design of educational assessments for such students.

The most important pieces of legislation in the United States in this area are the Individuals with Disabilities Education Act (IDEA) 1990, its predecessors (the 1969 Children with Specific Learning Disabilities Act, which was included in the Education of the Handicapped Act 1970, and the Education for All Handicapped Children Act 1975) and its subsequent reauthorizations, particularly those of 1997 and 2004. The 1997 revision, in particular, required all states that choose to accept IDEA funding (all, in fact, have) to arrange for the participation of students with disabilities in state and local accountability systems, with details of how each student will participate

and the support needed for such participation being specified in the student's Individualized Education Program (IEP).

The primary ways in which the participation of students with disabilities in large-scale testing programs has been arranged are through *test accommodations* and *alternative assessments*. Test accommodations are defined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) as variations in the specified procedures for test administration in response to a student's disability. The National Center on Educational Outcomes classifies these accommodations under the headings of (a) presentation, (b) equipment and materials, (c) response, (d) scheduling and timing, and (e) settings (National Center on Educational Outcomes, 2008).

The most common accommodations for test presentation are large-print versions of the test for students with visual impairments; versions of the test in Braille; the use of a scribe or sign interpreter; the opportunity for the student to have instructions repeated, reread, or clarified; and translation of the directions into the student's native language. Equipment and materials accommodations include magnification and amplification equipment, special acoustic conditions, and the use of a calculator. Accommodations for responding to the test items include the provision of a scribe, a computer keyboard, a brailler, or allowing the student to write rather than bubble in an item response. Most states also allow extended time to take the test, allowing the student to take breaks; scheduling the test over multiple sessions, or even over multiple days; and scheduling test times to suit the student. Finally, many states allow students with disabilities to take the tests in a room on their own, in a carrel, in a small group, or even outside the school (e.g., at home or in a hospital).

Where students with disabilities are not able to participate in mandated tests even with test accommodations, *alternative assessments* may be offered. As Pullin (2008) notes, these *alternative* assessments frequently address different standards from the tests taken by students without disabilities and can take very different forms, including out-of-grade testing and portfolios of work that are assessed either by reference to the grade-level expectations and achievement level descriptors specified for all students in that grade or relative to the student's IEP rather than to the relevant state standards (Pullin, 2008).

There is little doubt that accommodations and alternative assessments have in many cases had the effect "of altering the content of special education away from the low-level functional-life-skills approaches traditional for students with more severe disabilities in favor of more academic content associated with the curricular standards articulated for general education" (Pullin, 2008, p. 123). Furthermore, in states that link the award of high school diplomas to performance on standardized tests, such accommodations and alternative assessments have undoubtedly increased the number of students with disabilities who are able to receive high school diplomas (Ysseldyke, Dennison, & Nelson, 2003). However, although the efforts to allow students with disabilities to access a richer curriculum must be applauded, it must also be acknowledged that they take us into unfamiliar territory, particularly in terms of the arguments made in this chapter regarding the centrality of constructs.

It is clear that alternative assessments almost invariably assess different constructs from the assessments administered to students without disabilities, but it is also important to note that even apparently quite minor test accommodations can also affect the construct being assessed.

In the early development of tests, it was common to classify tests as either power tests or speeded tests, depending on whether the primary determinant of a student's score was the accuracy of the answers or the number of items completed. A common definition of speededness is that a test is (at least partially) speeded for a particular group of examinees if any of the examinees fail to complete three fourths of the items in the test and if less than 80% of the candidates complete all the items in the test (Davies, Kaiser, & Boone, 1987), although as Ellerin Rindler (1979) has pointed out, power and speededness interact in ways that are difficult to predict. It has been known for well over half a century that speeded and unspeeded versions of the same test may be measuring different constructs (Davidson & Carroll, 1945) so that even minor accommodations may change the construct addressed by an assessment.

The central argument of this chapter is that because construct interpretations are at the heart of validity argument, the design of assessments should begin by defining the construct to be assessed, and only then designing assessments that will yield evidence that support inferences regarding the construct, as is made manifest in the evidence-centered design paradigm (Mislevy, Almond, & Lukas, 2003). Where the process takes place in reverse, then validation becomes a values-based process rather than the technical process it should be, and in particular, the values of the test designer can feed into test design, resulting in a shift from making the important measurable to making the measurable important, and with test outcomes that are difficult, if not impossible, to interpret. If it is accepted that construct definition should precede assessment design for the general population, then it should also apply to students with disabilities. In other words, the debate should not be about assessments but constructs. The assessment of all students—with and without disabilities—should begin from a consideration of constructs, via a necessarily value-laden debate about whether the constructs should apply to all students or just to some. As the assessment is developed, it may be that accommodations are built in as optional procedures for administration, but these should be such that they do not change the construct being assessed. For example, the provision of large-print test booklets will improve the performance of some students with visual impairments but would not do so for students without visual impairments. As such, the validity of the test would be the same if all students were provided with large-print test booklets—the construct would not have changed—but considerations of efficiency would support the idea that the additional cost of such booklets should be borne only in the cases of students who would benefit from this provision. However, where accommodations would result in a change of construct being addressed, then the reason for this should be examined not in the assessment but in the appropriateness of the construct for the portion of the population in question. In this way, there is at least the prospect of treating all students equitably in terms of their assessment, irrespective of their (dis-)abilities.

CONCLUSION

Over the past century, the notion of validity has developed from a property of assessments, to a property of scores, and finally to a property of "inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). The idea that construct interpretations are at the heart of validity argument and that therefore construct definition is essential to effective assessment has also now become widely accepted, at least within the measurement community. Such an approach suggests that whereas contestations about the ability of a particular assessment procedure to support valid inferences *may* be related to technical issues about the design and implementation of the procedures, they are more likely to be the result of differences of view about the kinds of construct interpretations that are intended to be made from assessment outcomes. This is important because it separates, to an extent at least, matters of fact from those of value. The extent to which particular assessment procedures support particular kinds of inferences involves the integration of empirical evidence and theoretical rationales and is, therefore, principally, a technical matter. Whether the kinds of inferences that the assessment supports are the right inferences to be drawn is, to a much greater extent, a matter of value.

The argument of this chapter has been that the consequences of this hard-won consensus have not been effectively followed through to the social settings within which assessments are administered. Specifically, I have argued that separating the process of defining the construct to be assessed from the construction of the assessment to assess that construct provides a perspective that is useful in advancing equity.

In the case of assessing history, multiple-choice tests can support reasonably valid inferences about the extent to which examinees know facts and dates, but they are much less able to support inferences about the ability of examinees to weigh evidence and assemble historical arguments. The fact that boys do better on multiple-choice tests of history is often cited as evidence of the bias of multiple-choice tests against female examinees, but in this chapter, I have argued that it is more helpful to locate the inequity in the choice of a definition of construct of history in a way that has differential impact on males and females. In the same way, tests of mental rotation of three-dimensional solids are not biased against females because males really do tend to be better at this particular skill. We could try to find items of mental rotation that minimize the differences between males and females (in a similar way to the approach used by Freedle with the SAT), but surely a more appropriate response is to question whether the construct of mentally rotating three-dimensional solids is important (and if it is, what kinds of educational experiences are helpful in developing this skill).

In the case of admission to higher education, we could blame the SAT for its adverse impact on students of color, but this is likely to be less effective in addressing unequal access to higher education than looking at the experiences of students of color in K–12 education and what kinds of support such students would need to be successful in higher education. In the same way, I have suggested that a greater focus

on the construct, rather than the assessment, would frame the debate about the relative merits of performance assessment in a more helpful way.

Finally, in the assessment of students with special educational needs, I have argued that where test accommodations and alternative assessments change the construct of interest, it is essential that this is done through a focus on the construct rather than ad hoc modifications of the assessments, which lead to results that are difficult, if not impossible, to interpret. If a particular construct that is defined for one population is not suitable for another, then it seems to me to be much better to define a construct that *is* appropriate for that second population rather than simply modify the assessment.

In many cases, of course, adverse impact on female students, on students of color, or on students with special needs is the result of deficiencies in the assessment, which thus brings the validity of the assessment for its intended use into question. At other times the assessments will be supporting valid inferences about capabilities that are unequally distributed between different groups of students. Our chances of advancing the cause of equity in education will, I believe, be greatly enhanced if we know which is which.

ACKNOWLEDGMENTS

I would like to thank the developmental reviewers—Val Klenowski and Robert Rueda—and the volume editors—Allan Luke, Judith Green, and Greg Kelly—for helpful comments on earlier drafts of this chapter. Any deficiencies that remain, however, are, of course, entirely mine.

REFERENCES

Abrami, P. C., & Barrett, H. (2005). Directions for research and development on electronic portfolios. *Canadian Journal of Learning and Technology*, *31*(3), 1–15.

Access to Medicine. (2009). *Access to medicine*. Retrieved July 31, 2009, from http://www.accesstomedicine.org/

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (4th ed.). Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement*, *51*(2, Part 2), 1–38.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Psychological Association.

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, *92*(Summer), 2–5, 21.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19–32). Hillsdale, NJ: Lawrence Erlbaum.
Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*. London: McKinsey.
Bechtoldt, H. P. (1951). Selection. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1237–1267). New York: Wiley.
Bechtoldt, H. P. (1959). Construct validity: A critique. *The American Psychologist*, *14*, 619–629.
Belenky, M. F., Clinchy, B. M., Goldberger, N. R., & Tarule, J. M. (1986). *Women's ways of knowing: The development of self, voice, and mind*. New York: Basic Books.
Berlak, H. (1992). Toward the development of a new science of educational testing and assessment. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, et al. (Eds.), *Towards a new science of educational testing and assessment* (pp. 181–206). Albany: State University of New York Press.
Bishop, J. H. (2001a). A steeper, better road to graduation. *Education Next*, *1*(4), 56–61.
Bishop, J. H. (2001b). *Why do students learn more when achievement is examined externally?* Retrieved June 11, 2009, from http://media.hoover.org/documents/ednext20014unabridged_bishop.pdf
Block, J. H. (1976). Issues, problems and pitfalls in assessing sex differences: A critical review of The Psychology of Sex Differences. *Merrill-Palmer Quarterly*, *22*, 283–308.
Boaler, J. (1997). *Experiencing school mathematics: Teaching styles, sex and setting*. Buckingham, UK: Open University Press.
Boaler, J. (2008). Promoting "relational equity" and high mathematics achievement through an innovative mixed-ability approach. *British Educational Research Journal*, *34*, 167–194.
Bowlby, J. (1969). *Attachment and loss: Vol. 1. Attachment*. New York: Basic Books.
Brasel, K. J., Bragg, D., Simpson, D. E., & Weigelt, J. A. (2004). Meeting the Accreditation Council for Graduate Medical Education competencies using established residency training program assessment tools. *American Journal of Surgery*, *188*, 9–12.
Breland, H. M. (1991). *A study of sex differences* (Research Rep. No. RR-91-61). Princeton, NJ: Educational Testing Service.
Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Washington, DC: American Council on Education/Praeger.
Broome, E. C. (1903). *A historical and critical discussion of college admission requirements*. New York: Macmillan.
Burton, L. (2004). *Mathematicians as enquirers: Learning about learning mathematics*. Dordrecht, Netherlands: Kluwer Academic.
Calfee, R., Lau, E., & Sutter, L. (1983). Establishing instructional validity for minimum competency programs. In G. F. Madaus (Ed.), *The courts, validity and minimum competency testing* (pp. 95–113). Boston: Kluwer Academic.
Campbell, D. T. (1976). *Assessing the impact of planned social change*. Hanover, NH: Dartmouth College Public Affairs Center.
Cohen, L., Mannion, L., & Morrison, K. (2004). *A guide to teaching practice*. London: Routledge.
College Board. (2009). *New SATs for the press*. Retrieved July 31, 2009, from http://www.collegeboard.com/about/news_info/sat/faqs.html
Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform and rebellion*. Hillsdale, NJ: Ablex.
Council of Heads of Medical Schools. (1998). *Statement of principles*. London: Author.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
Crouse, J., & Trusheim, D. (1988). *The case against the SAT*. Chicago: Chicago University Press.
Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (2nd ed., pp. 621–694). Washington, DC: American Council on Education.

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time limit scores: A factor analysis. *Educational and Psychological Measurement*, *5*, 411–427.
Davies, T., Kaiser, R., & Boone, J. (1987). *Speededness of the Academic Assessment Placement Program (AAPP) reading comprehension test*. Nashville: Board of Regents of the State University and Community College System of Tennessee.
Department for Children, Schools and Families. (2008). *GCSE and equivalent results in England 2007/08 (provisional)*. London: Author.
Dorans, N. J. (1999). *Correspondence between ACT and SAT I scores* (Vol. RR 99-2). Princeton, NJ: Educational Testing Service.
Dorans, N. J. (2004). Freedle's table 2: Fact or fiction? *Harvard Educational Review*, *74*(1), 62–72.
Durán, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education*, *32*, 292–327.
Eckstein, M. A., & Noah, H. J. (1993). *Secondary school examinations*. New Haven, CT: Yale University Press.
Ellerin Rindler, S. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261–270.
Feingold, A. (1988). Cognitive gender differences are disappearing. *The American Psychologist*, *43*, 95–103.
Ferguson, R. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal of Legislation*, *28*, 465–498.
Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method of reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1–43.
Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, *59*, 185–213.
Gardner, H. (1989). Zero-based arts education: An introduction to Arts PROPEL. *Studies in Art Education: A Journal of Issues and Research*, *30*(2), 71–83.
Gardner, H. (1992). Assessment in context: The alternative to standardised testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 77–117). Boston: Kluwer Academic.
Garlick, P., & Brown, G. (2008, May 17). Widening participation in medicine. *British Medical Journal*, *336*, 1111–1113.
Garrett, H. E. (1937). *Statistics in psychology and education*. New York: Longmans, Green.
Gilligan, C. (1982). *In a different voice*. Cambridge, MA: Harvard University Press.
Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Cambridge, MA: Harvard University Press.
Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427–438.
Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Science*, *13*(2), 65–73.
Hanson, F. A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley: University of California Press.
Hart, K. M. (Ed.). (1981). *Children's understanding of mathematics: 11–16*. London: John Murray.
Hezlett, S. A., Kuncel, N. R., Vey, M. A., Ahart, A. M., Ones, D. S., Campbell, J., et al. (2001). *The effectiveness of the SAT in predicting success early and late in college: A comprehensive meta-analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum.
Hubin, D. R. (1988). *The Scholastic Aptitude Test: Its development and introduction, 1900–1948*. Unpublished doctoral dissertation, University of Oregon, Eugene.

Hume, D. (1896). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. Oxford, UK: Clarendon Press. (Original work published 1739)

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139–155.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494–495.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*, 8801–8807.

Jewitt, C. (2006). Multimodality and literacy in school classrooms. *Review of Research in Education*, *32*, 241–267.

Kellaghan, T. (Ed.). (1996). *Admission to higher education: Issues and practice*. Princeton, NJ: International Association for Educational Assessment.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book.

Kishiyama, M. M., Boyce, W. T., Knight, R. T., Jimenez, A. M., & Perry, L. M. (2009). Socioeconomic disparities affect prefrontal function in children. *Journal of Cognitive Neuroscience*, *21*, 1106–1115.

Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2002). *The utility of the SAT-I and SAT-II for admissions decisions in California and the nation* (Report No. 2002-6). New York: College Board.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D. M., Stecher, B. M., Klein, S. P., McCaffrey, D., & Deibert, E. (1994). *Can portfolios assess student performance and influence instruction? The 1991–92 Vermont experience* (Vol. RP-259). Santa Monica, CA: RAND Corporation.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus & Giroux.

Levine, D. O. (1986). *The American college and the culture of aspiration 1915–1940*. Ithaca, NY: Cornell University Press.

Lieberman, A., & Wood, D. R. (2002). *Inside the National Writing Project: Connecting network learning and classroom teaching*. New York: Teachers College Press.

Lindquist, E. F. (Ed.). (1951). *Educational measurement* (1st ed.). Washington, DC: American Council on Education.

Linn, M. C. (1992). Gender differences in educational achievement. In Educational Testing Service (Ed.), *Sex equity in educational opportunity, achievement, and testing: Proceedings of a 1991 ETS invitational conference* (pp. 11–50). Princeton, NJ: Educational Testing Service.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). Washington, DC: American Council on Education/Macmillan.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(Suppl. 9), 635–694.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Mathews, J. (2004). Whatever happened to portfolio assessment? *Education Next*, *4*(3), 73–75.

Maxwell, G. S. (2004, March). *Progressive assessment for learning and certification: Some lessons from school-based assessment in Queensland*. Paper presented at the third conference of the Association of Commonwealth Examination and Assessment Boards, Nadi, Fiji. Brisbane, Australia: University of Queensland Graduate School of Education.

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 625–652). Mahwah, NJ: Lawrence Erlbaum.

McGovern, C. (1994). *The SCAA Review of National Curriculum History: A minority report*. York, UK: Campaign for Real Education.

Messick, S. (1980). Test validity and the ethics of assessment. *The American Psychologist*, *35*, 1012–1027.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education/Macmillan.

Miles, E. (1998). *The Bangor dyslexia teaching system*. London: Whurr.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence centered design* (ETS Research Rep. No. RR-03-16). Princeton, NJ: Educational Testing Service.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, *30*, 109–162.

National Center on Educational Outcomes. (2008). *Participation and accommodation summary reports 2006–2007*. Retrieved June 28, 2009, from http://data.nceo.info/pa-summaries.asp

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago: Consortium on Chicago School Research.

Nuttall, D. L. (1987). The validity of assessments. *European Journal of Psychology of Education*, *2*, 109–118.

O'Connell, A. N. (1990). Eleanor Emmons Maccoby. In A. N. O'Connell & N. F. Russo (Eds.), *Women in psychology: A bio-bibliographic sourcebook* (pp. 231–237). New York: Greenwood.

O'Day, J., & Smith, M. S. (1993). Systemic school reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 250–312). San Francisco: Jossey-Bass.

Organisation for Economic Cooperation and Development. (2008). *Education at a glance 2008*. Paris: Author.

Piché, D. M., & Taylor, W. L. (Eds.). (1991). *A report on shortchanging children: The impact of fiscal inequity on the education of students at risk* (Prepared for the Committee on Education and Labor, U.S. House of Representatives, One Hundred First Congress, Second Session). Washington, DC: Government Printing Office.

Pirie, S. E. B. (1987). *Nurses and mathematics: Deficiencies in basic mathematical skills among nurses—Development and evaluation of methods of detection and treatment*. London: Royal College of Nursing/Scutari Press.

Project Zero. (2005). *Arts PROPEL*. Retrieved July 31, 2009, from http://pzweb.harvard.edu/research/propel.htm

Pullin, D. (2008). Individualizing assessment and opportunity to learn: Lessons from the education of students with disabilities. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 109–135). Cambridge, UK: Cambridge University Press.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Shayer, M., & Adey, P. S. (1981). *Towards a science of science teaching: Cognitive development and curriculum demand*. London: Heinemann.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Vinner, S. (1997). From intuition to inhibition—Mathematics, education and other endangered species. In E. Pehkonen (Ed.), *Proceedings of the 21st conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 63–78). Lahti, Finland: University of Helsinki, Lahti Research and Training Centre.

Vogt, W. P. (1999). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. London: Falmer Press.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270.

Wiley, D. E. (2001). Validity of constructs versus construct validity. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 207–227). Mahwah, NJ: Lawrence Erlbaum.

Wiliam, D. (2001). Reliability, validity and all that jazz. *Education 3-13*, *29*(3), 17–21.

Wiliam, D. (2003). Constructing difference: Assessment in mathematics education. In L. Burton (Ed.), *Which way social justice in mathematics education?* (pp. 189–207). Westport, CT: Praeger Press.

Wiliam, D. (under review). Standardized student assessment in an age of accountability. *Educational Psychologist*.

Wiliam, D., Brown, M., Kerslake, D., Martin, S., & Neill, H. (1999). The transition from GCSE to A-level in mathematics: A preliminary study. *Advances in Mathematics Education*, *1*, 41–56.

Wiliam, D., Millar, M., & Bartholomew, H. (2004). *Selection for medical education: A review of the literature*. Retrieved August 1, 2008, from www.dylanwiliam.net

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

Wineburg, S. S., & Fournier, J. (1994). Contextualized thinking in history. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 285–308). Hillsdale, NJ: Lawrence Erlbaum.

Ysseldyke, J., Dennison, A., & Nelson, R. (2003). *Large-scale assessment and accountability systems: Positive consequences for students with disabilities* (Synthesis Rep. No. 51). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved June 28, 2009, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis51.html

Zenderland, L. (1998). *Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing*. Cambridge, UK: Cambridge University Press.